UNIVERZITA KARLOVA V PRAZE
FILOZOFICKÁ FAKULTA

DIPLOMOVÁ PRÁCE

2012 Eliška Rudá

UNIVERZITA KARLOVA V PRAZE

Filozofická fakulta

Katedra sociologie

Diplomová práce

Ing. Bc. Eliška Rudá

Teorie odpovědi na položku a její aplikace v sociologii

Item Response Theory and Its Application in Sociology

Praha 2012

Vedoucí práce: PhDr. Jiří Vinopal, Ph.D.

Prohlášení

Prohlašuji, že jsem diplomovou práci na téma „Teorie odpovědi na položku a její aplikace v sociologii“ vypracovala samostatně. Veškerou použitou literaturu a podkladové materiály uvádím v přiloženém seznamu literatury.

V Praze dne ……..

Podpis

Ráda bych tímto poděkovala svému vedoucímu diplomové práce PhDr. Jiřímu Vinopalovi, Ph.D. za projevenou vstřícnost a cenné připomínky. Zvláštní díky patří mé rodině, která mě během celých studií vroucně podporovala a osvobozovala mě od nejrůznějších domácích prací. Na tomto místě je také vhodné ocenit existenci nejrůznějších online komunit okolo tématu Item Response Theory, jejíž členové mi často dobrými tipy ušetřili mnoho času.

Teorie odpovědi na položku a její aplikace v sociologii

Abstrakt

Diplomová práce se věnuje modernímu metodologickému přístupu, teorii odpovědi na položku, která modeluje vztah latentní proměnné a měřitelné odpovědi na položku, v kontextu sociologie. Teoretický úvod osvětluje podstatu teorie a srovnává ji s klasickou teorií testů. Dále je nastíněno možné využití a aplikace teorie v sociologii. Práce je na závěr doplněna o ilustrativní příklad, který na reálných datech hodnotí nástroj na měření subjektivní kvality pracovního života právě z pohledu teorie odpovědi na položku.

Klíčová slova: teorie odpovědi na položku, škála, reliabilita, model, subjektivní kvalita pracovního života

The Item response Theory and its Application in the Sociology

Abstract

The focus of this master's thesis is an item response theory, a modern methodological approach, which models a relationship between a latent variable and measured responses to the particular item. A theoretical introduction, which explains fundamentals of the theory, is followed by a comparison with the classical test theory and suggestions of its possible applications in a field of sociology. The thesis is completed with an illustrative example using real data to evaluate an instrument for measuring subjective quality of working life in the framework of item response theory.

Keywords: Item response theory, scale, reliability, model, subjective quality of working life

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM ZKRATEK A SYMBOLŮ

1PL	One Parameter Logistic Model / Logistický model s jedním parametrem
2PL	Two Parameters Logistic Model / Logistický model se dvěma parametry
3PL	Three Parameters Logistic Model / Logistický model se třemi parametry
CTT	Classical Test Theory / Klasická teorie testů
DIF	Differential Item Functioning / Odlišné fungování položek (bias)
GRM	Graded Response Model / Model stupňovaných odpovědí
ICC	Item Characteristic Curve / Charakteristická křivka položky
IIC	Item Information Curve / Informační křivka položky
IRT	Item Response Theory / Teorie odpovědi na položku
MFS	Levin's Multilinear Formula Scoring
NCM	Nominal Category Model / Model pro nominální kategorie
RMSEA	Root Mean Square Error of Approximation
SEM	Standard Error of the Mean / směrodatná odchylka
TCC	Total Chracteristic Cruve / Charakteristická křivka testu
TIC	Total Information Curve / Informační přínos testu

Symboly

θ	Theta / latentní proměnná
a	parametr *a*
b	parametr *b*
c	parametr *c*
i	položka *i*
e	Eulerovo číslo

Úvod

Teorie odpovědi na položku (Item Response Theory - IRT) je moderním metodologickým přístupem, který se používá v mnoha disciplínách zahrnující i sociologii. Jedná se o matematické modelování vztahu latentní proměnné a měřitelné odpovědi na konkrétní položku v rámci testu. V České republice je teorie odpovědi na položku zatím velmi málo rozšířeným přístupem. Pokud je autorce známo, v kontextu české sociologie se jedná o téma málo zpracované (názor potvrzují i Urbánek a Šimeček 2001), ačkoliv IRT zásadně mění pohled na konstrukci dotazníků a vůbec na reliabilitu výzkumu.

Konkrétněji teorie odpovědi na položku přispívá ke zvyšování kvality sociologických výzkumů tím, že na základě modelů umožňuje uspokojivě zodpovídat otázky funkce každé položky v dotazníku, její rozlišovací schopnosti a informačního přínosu pro test. IRT také pomáhá sestavovat krátké, cílené a reliabilní škály s optimálním počtem kategorií. To, že IRT zkoumá každou položku dotazníku odděleně, dává prostor pro srovnávání sociologických výzkumů při použití odlišných technik dotazování či při změnách formátu otázky či délky dotazníku, případně pomáhá odhalit tzv. bias.

Diplomová práce si klade za cíl vystihnout podstatu teorie odpovědi na položku a následně uvedenou teorii využít na sekundárních datech a ukázat její možnosti pro sociologii. Výstupem práce bude formulace sociologicky relevantní podstaty teorie odpovědi na položku a přehled, jak může přispět k vyšší kvalitě sociologického výzkumu. Možnosti aplikace teorie odpovědi na položku v sociologii budou

demonstrovány jak na úrovni teoretické, tak na úrovni konkrétního komplexního ilustrativního příkladu na reálných datech.

Práce je rozdělena kromě úvodu a závěru do třech částí. První teoretická kapitola osvětluje podstatu teorie odpovědi na položku a představuje stručně základní modely, předpoklady a omezení použité metodologie širší odborné veřejnosti. Dále následuje srovnání IRT s klasickou teorií testů, její výhody a nevýhody. V druhé kapitole jsou nastíněny možnosti využití teorie odpovědi na položku v sociologii i s konkrétními příklady a diskutován aktuální stav využití ve výzkumu v České republice. Třetí kapitola demonstruje využití IRT na datech subjektivního hodnocení spokojenosti s pracovním životem (Vinopal 2011).

Výstupy diplomové práce mohou být využity na úrovni praktické ve formě optimalizace a doplnění opakujícího se výzkumu subjektivní spokojenosti s pracovním životem (Vinopal 2011). Na práci lze také pohlížet jako na námět pro další výzkumy, pro které je využití teorie odpovědi na položku relevantní. Vzhledem k originalitě přístupu v rámci české sociologie přispěje diplomová práce i ke zvýšení informovanosti akademické obce o problematice a hlavně možnostech využití a přínosech IRT.

1. Teoretický exkurz

Na teoretickou část této práce je kladen velký důraz, neboť povaha tématu si jej vyžaduje. Teorie odpovědi na položku je citlivá na zvolení vhodného modelu se správným počtem dobře odhadnutých parametrů. Teorie je vždy doplněna o ukázkové grafické výstupy z reálných dat (Vinopal 2011), která budou pak komplexně zpracována v ilustrativním příkladu (viz kapitola 3). Dalším z důvodů je až na výjimku (Jelínek, Květoň & Vobořil 2011) nedostatek česky psané literaturu k tématu teorie odpovědi na položku, navíc se sociologickým zaměřením.

1. 1. Historie a podstata IRT

Ačkoliv 20. století bylo ve znamení klasické teorie testů, dá se historie teorie odpovědi na položku vysledovat ve dvou relativně nezávislých hlavních proudech (Embertson & Reise 2000). V americké větvi se jako milník považuje vydání knihy *Statistical Theories of Mental Test Scores* autory Lord a Novick (1968). Principy teorie odpovědi na položku byly již zmiňovány a částečně používány dříve, nicméně tato kniha poskytla první ucelené a konzistentní zpracování tématu. Zároveň byla impulsem pro další využívání a systematický rozvoj oblasti ať už v univerzitním prostředí či přímo v praxi při psychometrických výzkumech (Embertson & Reise 2000: 5).

Druhá, evropská větev, začíná v šedesátých letech u dánského matematika Georga Rasche, který se věnoval vývoji testů pro dánskou armádu (Embertson & Reise 2000: 5). Takzvaný „Raschův model“ (Rasch 1981), ke kterému položil základy, je matematicky ekvivalentní jednoparametrovému logistickému modelu (více viz kapitola

1. 3. 2). Nicméně uživatelé tohoto poměrně rozšířeného modelu nepoužívají pro své výzkumy název IRT (DeMars 2010: 15).

V současné době se již oba vývojové proudy IRT spojily do jednoho a teorie odpovědi na položku se bouřlivě rozvíjí (Embertson & Reise 2000: 13). A to jak na úrovni metodologie, kdy vznikají nové software pro výpočty (např. IRTPRO, flexMIRT) či nové modely[1], tak na úrovni využívání. Stále významnější místo zaujímá teorie odpovědi na položku především v edukativních a psychometrických testech (Embertson & Reise 2000: 13).

Podstata teorie odpovědi na položku (IRT) se však nemění. Matematické modely vniklé na základě IRT znázorňují vztah mezi měřenou latentní proměnnou a odpovědí na konkrétní položku. Latentní proměnnou, která je v literatuře a i v této diplomové práci běžně značená jako θ [theta], může být dovednost, charakterová vlastnost, talent, znalosti, ale i například pro sociology zajímavější postoje (DeMars 2010: 9). Položkou značíme jednu konkrétní otázku, respektive odpověď na ní. Je lépe se vyvarovat pojmu otázka, neboť jedna otázka v dotazníku může obsahovat několik položek, jako například baterie.

IRT modely jsou založeny na pravděpodobnosti, s jakou určitý respondent odpoví na konkrétní položku v souladu s jeho úrovní měřené latentní proměnné θ. Tím se podobají logistické regresi (Reeve & Fayers 2005: 57) a vyžadují sofistikovanější matematický aparát než klasické testy (CTT), které spočívají především na součtech a průměrech.

[1] Například „double monotonicity model“ vypracovaný I. W. Molenaar (1997)

1. 2. Parametry

Různé typy modelů vytvořené v rámci teorie odpovědi na položku obsahují jeden až tři parametry[2]. Každý z parametrů modelu se nejčastěji odhaduje na základě empirických dat (více viz kapitola 1. 4. 3) a každému náleží patřičná interpretace.

Parametr *a*, schopnost diskriminace položky[3], je užitečný pro výběr položek, které dobře rozlišují mezi respondenty s vysokou a nízkou mírou měřené latentní proměnné, mezi respondenty, kteří znají a neznají testovaný předmět, či mezi těmi, kteří mají negativní a pozitivní postoj. Vysoká diskriminace položek je žádoucí, neboť několik dobře zvolených otázek umožní změřit latentní proměnnou lépe než mnoho špatně rozlišujících otázek.

V CTT najdeme obdobu parametru *a* například v bodově biserálním koeficientu (DeMars 2010: 6). V IRT je parametr *a* někdy také označován jako „sklon“. Při vyjádření pravděpodobnosti žádoucí odpovědi a latentní proměnné (viz Graf 3) můžeme parametr *a* odečíst právě ze strmosti, sklonu funkce. Čím má křivka větší sklon, tím je větší parametr *a*, a tím lépe daná položka diskriminuje.

Parametr *b*, obtížnost položky, slouží pro složení vhodného dotazníku cílové skupině. Jiný bude test z matematiky pro první a pro pátou třídu na základní škole. Pokud by dostali žáci v pátém ročníku příliš lehký test, průměrný student by odpověděl na téměř všechny otázky správně a jeho znalosti matematiky by zůstaly nezměřeny. Obdobně se dá očekávat jiný dotazník měřící postoje k násilí v běžné populaci či ve skupině delikventů sedící za mřížemi.

---

[2] Existuje i čtvrtý parametr *d*, parametr „ledabylosti“ (carelessness), který se však v testování prakticky nepoužívá, proto nebude zahrnut ani do této diplomové práce.

[3] Parametr *a*, schopnost diskriminace položky, bývá také některými autory (Urbánek a Šimeček 2001) překládán a nazýván jako „rozlišovací účinnost“.

Obtížnost položky nalezneme i v klasické teorii testů (CTT), jako podíl respondentů, kteří položku zodpověděli správně (popř. v určitém směru při měření postojů). Oproti tomu parametr *b*, obtížnost položky, v teorii odpovědi na položku (IRT) vyjadřuje pravděpodobnost, se kterou na danou položku odpoví správně (či v určitém směru) 50 % respondentů (DeMars 2010: 6).

Třetí **parametr *c*, uhádnutelnost**, umožňuje postihnout situace, kdy velkou roli hraje dimenze uhádnutelnosti, typicky při tzv. multiple choice testech[4]. V sociologii jde častěji o případy, kdy jsou některé odpovědi systematicky pod/nadhodnocovány, protože jsou sociálně ne/žádoucí (sociální desirabilita). Parametr *c* je vlastně pravděpodobnost správné odpovědi na položku *i,* i když osoba odpověď nezná, či pro sociology relevantnější situace, kdy respondent odpovídá v sociálně žádoucím směru, ačkoliv realita je jiná.

Sociální desirabilita úzce souvisí s tématem validity a v sociologických výzkumech je řešena už při sběru dat, na což existují různé techniky, ať už způsob sběru dat, „detektory lži“, technika náhodné odpovědi a další (více viz Chylíková 2011). V rámci klasické teorie testů tedy neexistuje jedinečná obdoba parametru *c*. Přehled ostatních parametrů a jejich alternativ je shrnut v Tab. 1. Celkové srovnání teorie odpovědi na položku a klasické teorie testů bude následovat v kapitole 1. 6.

Tab. 1 Vyjádření parametrů v IRT a jejich obdoba v CTT - shrnutí

Parametr	**Klasická teorie testů (CTT)**	**Teorie odpovědi na položku (IRT)**
a - diskriminace	Bodově biserálním koeficient	Sklon křivky
b - obtížnost	Podíl respondentů, kteří položku zodpovědí žádoucím způsobem. Průměr.	Pravděpodobnost, se kterou na položku odpoví žádoucím způsobem 50 % respondentů
c - uhádnutelnost		Asymptota

Zdroj: vlastní zpracování podle DeMars 2010

[4] Multiple choice testy jsou otázky, ve kterých je na výběr až několik správných a několik špatných odpovědí. Každá odpověď pak tvoří jednu dichotomickou položku.

1. 3. Modely

Tato podkapitola má poskytnout stručný, nikoliv však vyčerpávající přehled modelů používaných v rámci IRT. Důraz je kladen spíše na použitelnost a aplikovatelnost modelů s vysvětlením jejich podstaty, než na matematickou akurátnost a vyjadřování. Základní modely vycházejí z publikace DeMars 2010.

1. 3. 1. Datový soubor a technické informace

Veškeré modely, které následují, budou prakticky ukázány a vypočítány na datovém souboru, který pochází z výběrového šetření „Naše společnost“. Byl realizován pod názvem Stres na pracovišti – možnosti prevence v roce 2009 Centrem pro výzkum veřejného mínění, SOÚ AV ČR, v. v. i. Na základě kvótního výběru bylo vybráno 950 zaměstnanců v České republice ve věku od 18 do 65 let. Celkově bylo dotázáno 836 respondentů metodou standardizovaného rozhovoru s tazatelem na základě dotazníku.

Z výzkumu byla vybrána baterie osmnácti otázek (číslo 34, viz Příloha 1) měřící subjektivní spokojenost s pracovním životem na šestibodové škále. Zaznamenané odpovědi „neví“ a „netýká se“ byly brány jako chybějící odpovědi a nebyly zahrnuty.

Škála z datového souboru pro modelaci a ilustraci dichotomních modelů byla dichotomizována na spokojen a nespokojen. Autorka si je vědoma úskalí toho sloučení. Nicméně jde pouze o modelování jednotlivých variant bez cíle je jakkoliv interpretovat a činit z nich závěry. Stejně tak by pro tento účel mohla být použita jakákoliv jiná data.

Veškeré výpočty v rámci teorie odpovědi na položku byly zpracovány v programu IRTPRO (Item Response Theory for Patient-Reported Outcomes). Pro finální úpravy grafů byl použit program Adobe Illustrator.

1. 3. 2. Modely pro dichotomní položky

Dichotomní položky mají jen dvě kategorie odpovědí, které se standardně kódují 1 a 0[5]. Ve znalostních testech jsou správné odpovědi běžně označované 1, u měření např. postojů se číslem 1 označuje taková odpověď, která znamená vyšší úroveň měřeného konstruktu. Když například při měření spokojenosti s kvalitou pracovního života, odpověď „ano, souhlasím" značí vyšší spokojenost, bude označena 1. Pokud odpověď „nesouhlasím" indikuje vyšší spokojenost, bude taktéž označena 1. Pro čitelnost textu bude každá taková „jedničková" odpověď označovaná jako „žádoucí".

Mezi dichotomní modely se zařazují i položky s vícenásobnou volbou, která je však pro účely vyhodnocování převedena do dvou kategorií; typickým příkladem je správná/špatná odpověď (Jelínek, Květoň & Vobořil 2011: 35) v multiple choice testech.

Jednoparametrový logistický model (1PL)

Jednoparametrový logistický model je nejjednodušší z modelů, neboť vztah mezi úrovní latentní proměnné θ a žádoucí odpovědi je dán pouze jedním parametrem b, tedy obtížností položky i. Charakteristická křivka položky (ICC, Item Characteristic Curve) vyjadřuje pravděpodobnostní funkci položky. U jednoparametrového logistického modelu je vyjádřena vzorcem:

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}}$$

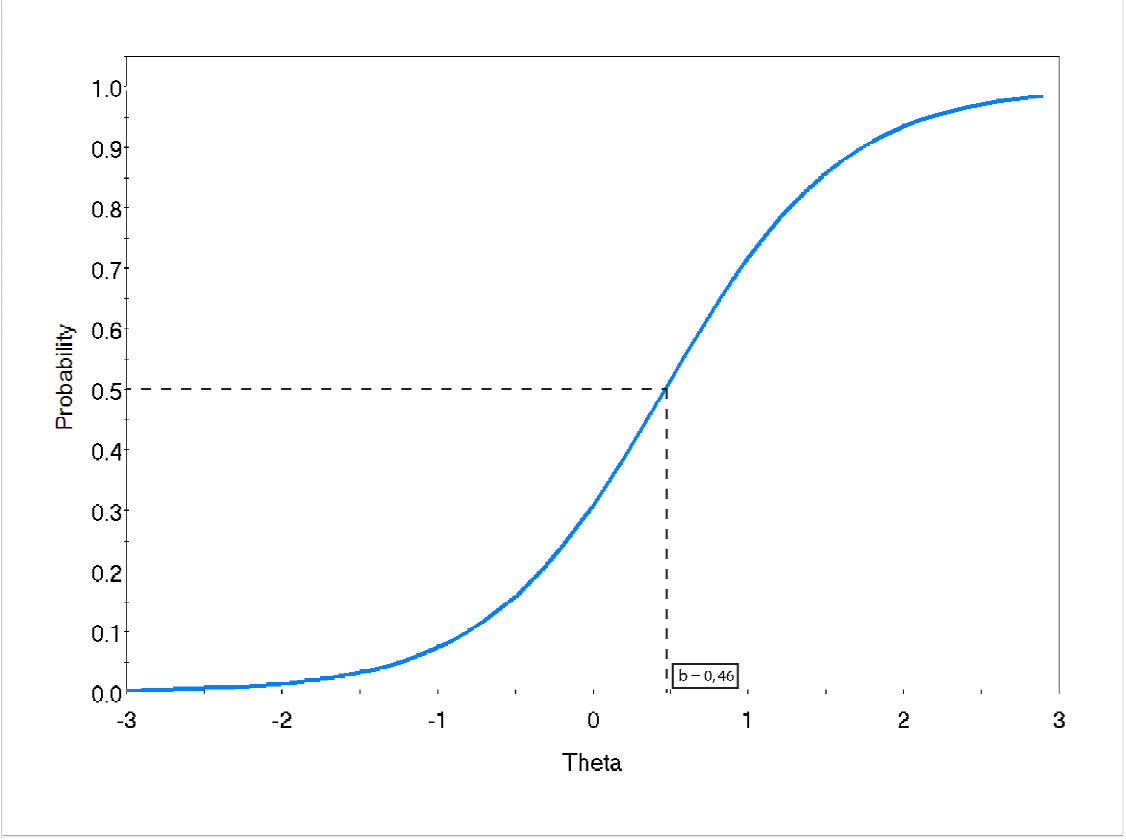
kde e je tzv. Eulerovo číslo, zaokrouhleně 2,718, a b_i jediný parametr modelu.

Funkce je monotónně rostoucí (viz Graf 1), tedy s vyšší úrovní latentního rysu se zvyšuje i pravděpodobnost odpovědi na položku v požadovaném směru. V případě otázky 34a (Jak jste spokojen s výší platu nebo mzdy) je parametr b roven 0,46.

[5] Pro použití v programu IRTPRO (více viz kapitola 3. 1. 2) je standardně třeba překódovat odpovědi na 1 (ano) a 2 (ne).

Parametr *b* lze získat odečtením hodnoty z grafu, pokud se pravděpodobnost rovná 0,5, tedy přesně polovina respondentů odpoví v požadovaném směru.

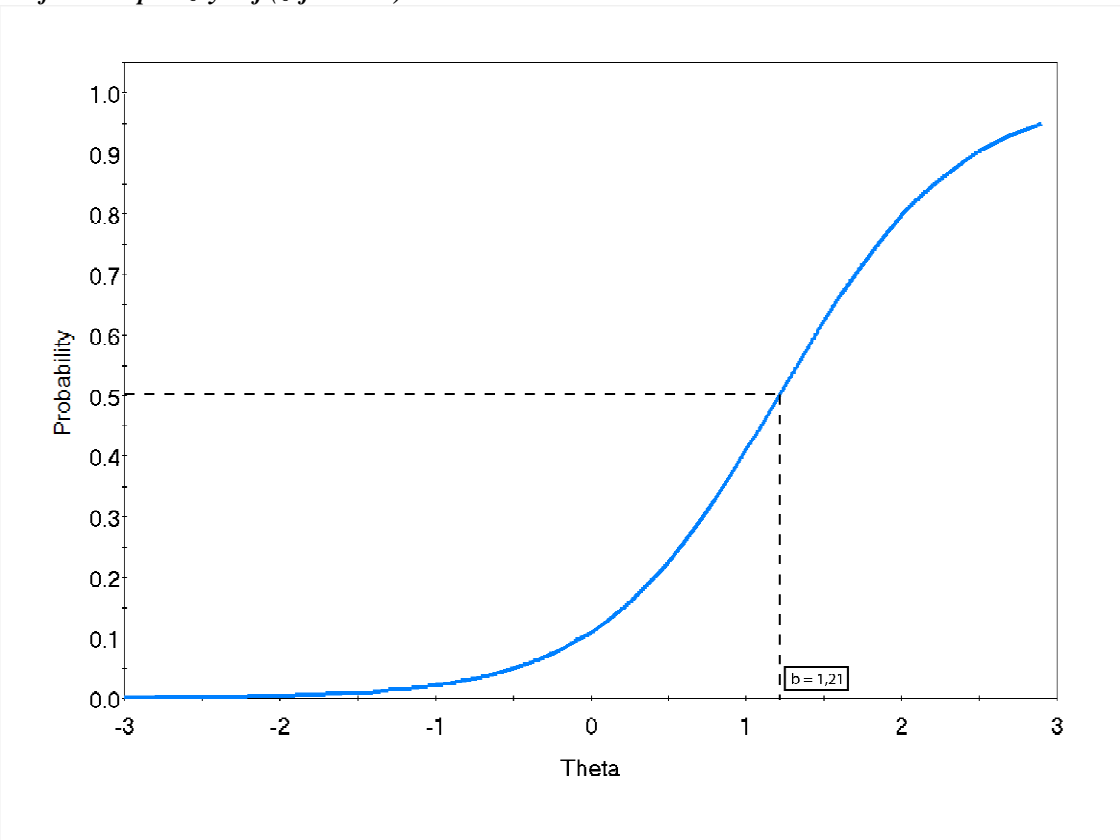
Graf 1: ICC dichotomizované položky 34a (plat) dle 1PL modelu



Zdroj: vlastní výpočty v IRTPRO

Obecně platí, že čím je parametr b vyšší, tím je položka „obtížnější“. V takovém případě má charakteristická křivka u logistického jednoparametrového modelu stále stejný tvar, pouze se „posune doprava“. Názorně tak můžeme porovnat křivku spokojenosti s platem (Graf 1, kde $b = 0{,}46$) a křivku spokojenosti se zajímavostí práce (viz Graf 2, kde $b = 1{,}21$). Interpretace by pak zněla, že položka dotazující se na zajímavost práce je „těžší“, neboť má větší parametr *b*, takže respondent musí být s prací více spokojen, aby odpověděl kladně i na tuto položku, než je tomu třeba u spokojenosti s platem.

Graf 2: ICC položky 34j (zajímavost) dle 1PL modelu



Zdroj: vlastní výpočty v IRTPRO

Matematickým ekvivalentem 1PL modelu je tzv. Raschův model, který sice byl vyvíjen samostatně, ale má obdobné znaky. Ačkoliv matematicky se Raschův model a jednoparametrický logistický model pod hlavičkou IRT neliší, je zde zásadní rozdíl v přístupu k empirickým datům. IRT se snaží vyvinout co nejlepší model, který odpovídá nasbíraným datům. Oproti tomu Raschův model je základ, do kterého se musí empirická data „napasovat". Pokud některé položky nebo respondenti nevyhovují modelu, jsou z další analýzy vyloučeny. Cílem rodiny Raschových modelů je získat jednoduchou a názornou interpretaci analýzy (Reeve 2002: 13).

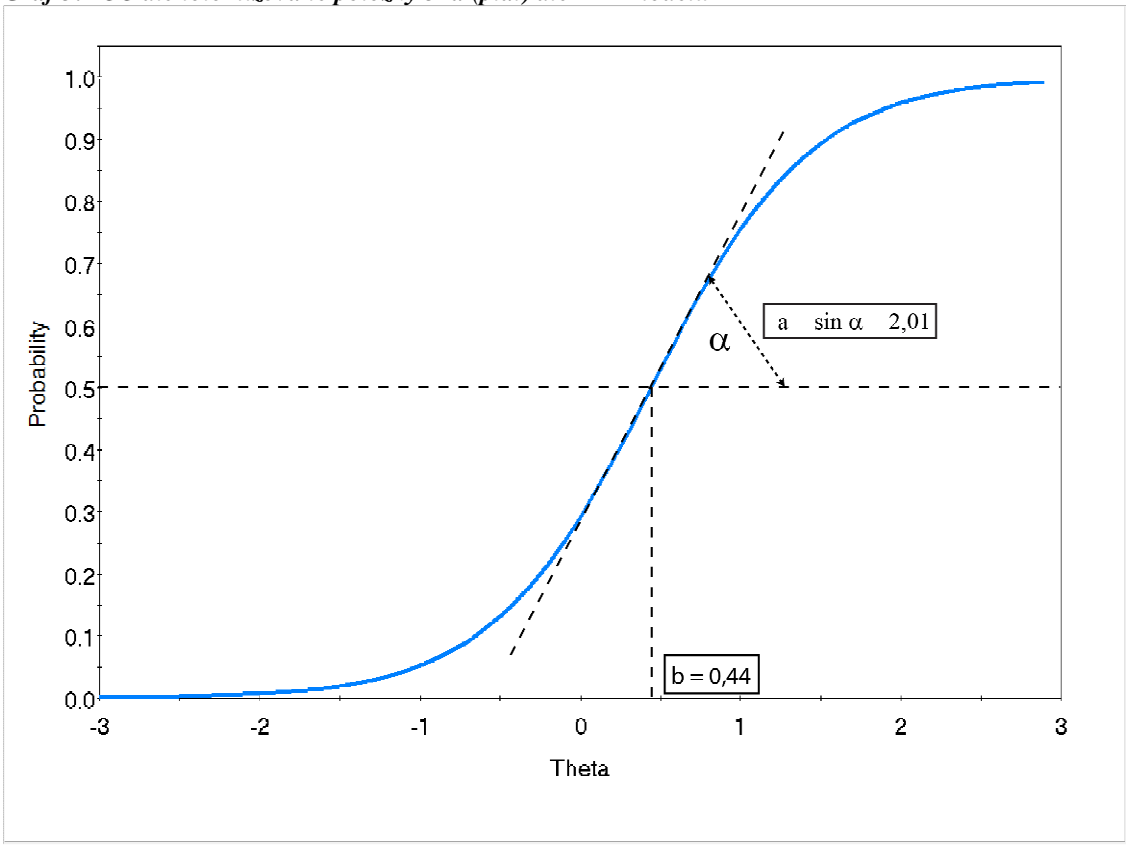
Dvouparametrový logistický model (2PL)

Dvouparametrový logistický model zahrnuje oproti jednoparametrovému modelu také vlastnost schopnost diskriminace položky, tedy parametr *a*. Model se tak blíží více realitě, neboť předpokládá, že položky souvisejí s latentní proměnnou v různé míře. Formálně lze model vyjádřit jako:

$$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

Z grafu lze parametr *a* odečíst jako sklon křivky ICC u pravděpodobnosti 0,5 (viz Graf 3). Čím strmější, tím lépe položka rozlišuje mezi respondenty.

Graf 3: ICC dichotomizované položky 34a (plat) dle 2PL modelu



Zdroj: vlastní výpočty v IRTPRO

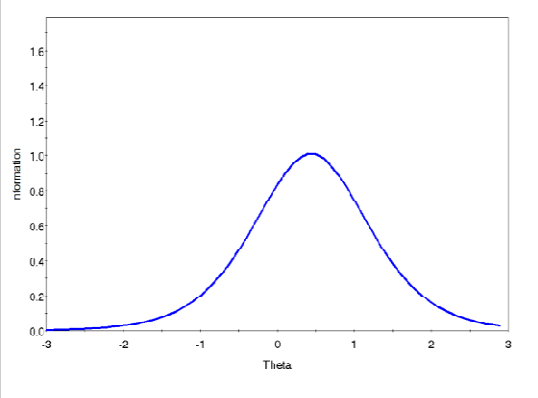
Kromě charakteristické funkce položky (ICC) lze zobrazovat i informační funkci položky (IIC). Ta ukazuje informační přínos každé položky pro danou úroveň latentního rysu θ. Na informační přínos položky lze pohlížet ze dvou hledisek. Jednak nás zajímá, pro (1) jakou hodnotu latentního rysu položka dobře rozlišuje; při jaké úrovni latentního rysu θ se funkce IIC maximalizuje. Konkrétně na otázce 34 a – spokojenost s výší platu (viz Graf 4) nás může zajímat, zda tato položka dobře rozlišuje u osob, které jsou celkově spokojeny či spíše nespokojeni s kvalitou pracovního života. ICC spokojenosti s výší platu dosahuje své maximální hodnoty při hodnotě θ = 0,4, tedy informační

přínos položky se maximalizuje spíše na pravé straně kontinua θ. Obdobná situace platí i pro položku 34 g (spokojenost s časovou náročností práce, viz Graf 5)[6].

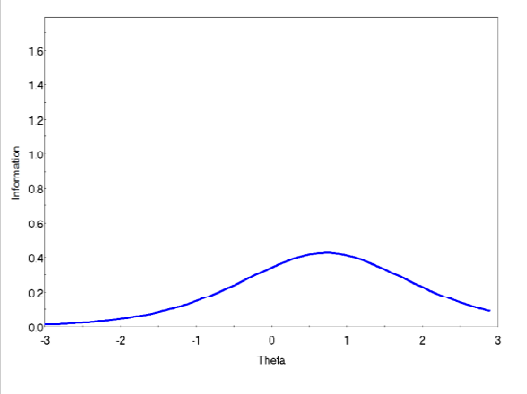
Dalším hlediskem (2) je právě schopnost diskriminace položky mezi danými respondenty s různou úrovní latentního rysu. Dle parametru *a* platí, že čím „strmější" křivka, tím lépe daná položka rozlišuje mezi respondenty (pro srovnání Graf 4 a Graf 5). Je vhodné zvážit, zda položky s nízkou diskriminační schopností z dotazníku nevyloučit úplně.

[6] Opět zde dochází k pouze názornému vysvětlení modelu, jeho parametrů a možnému využití. Nejedná se o interpretaci a relevantní posuzování dat z dotazníků, neboť není pro tento model (2PL) ověřena jeho nejlepší shoda s daty, která jsou navíc dichotomizována.

Graf 4: IIC dichotomizované položky 34a (plat) dle 2PL modelu



Graf 5: IIC dichotomizované položky 34g (časová náročnost) dle 2PL modelu



Zdroj: vlastní výpočty v IRTPRO

Tříparametrový logistický model (3PL)

Tříparametrový model představuje obecnější variantu, ze které lze odvodit i předchozí modely (1PL, 2PL) zafixováním parametrů. Kromě parametrů *a* a *b* se v modelu objevuje i parametr *c*, označovaný jako „uhádnutelnost". Tento model byl vyvinut pro znalostní testy tak, aby mohla být teorie odpovědi na položku použita i pro tzv. multiple choice testy, kde hádání hraje podstatnou roli (Reeves 2002: 19).

Pro sociologické výzkumy by se pravděpodobně hodilo jiné označení, protože při měření např. postojů nelze hovořit o špatné odpovědi, či o hádání správné odpovědi. Přesto tento parametr má i pro sociology smysl, například při problematice „sociálně žádoucí" odpovědi, či při systematickém nad/podhodnocováním odpovědí na některé otázky určitými skupinami respondentů. Matematické vyjádření modelu:

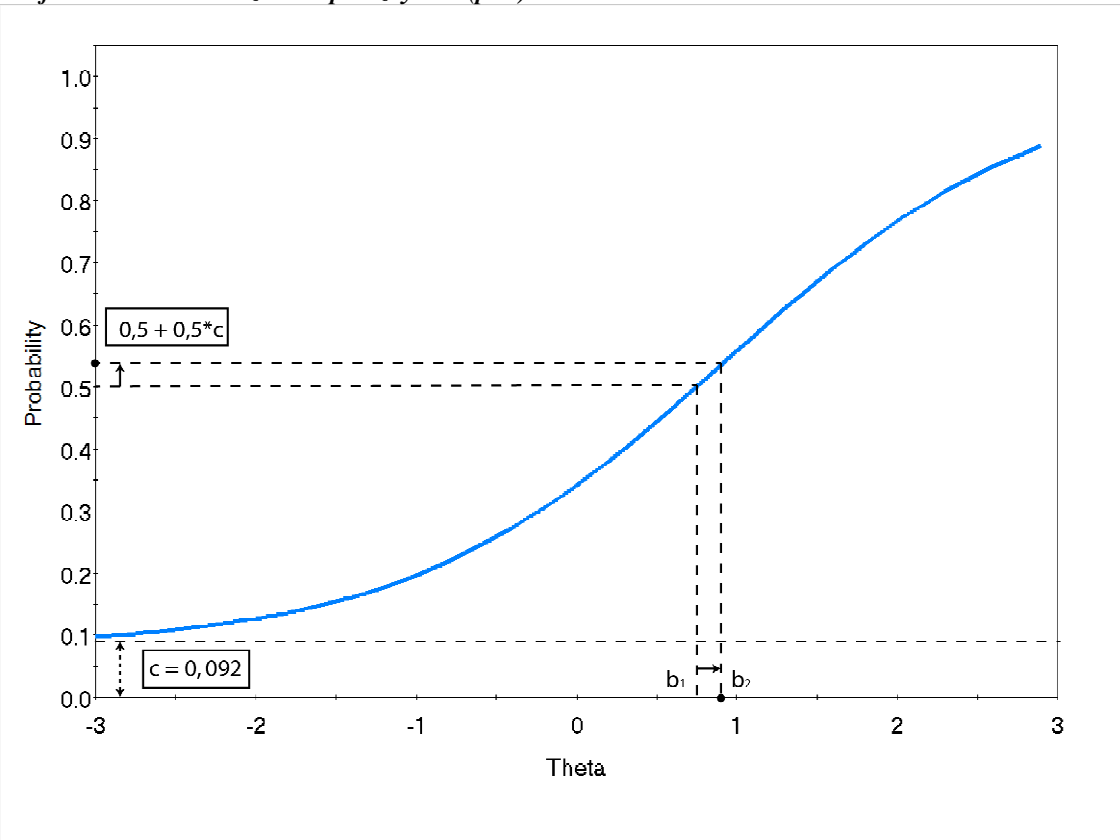
$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i)}}$$

Parametr *c* lze v grafu chápat jako „asymptotu"[7] charakteristické křivky položky (viz Graf 6). S větším parametrem *c* se zvyšuje podíl složky „uhádnutelnosti" v položce a vizuálně se i celý graf posouvá po ose y.

[7] Laicky řečeno asymptota je přímka, ke které se „blíží" křivka funkce, ale nikdy jí neprotne (limita funkce).

Přidání parametru *c* také upravuje hodnoty a interpretace dalších parametrů v modelu. Parametr *b* se mění z hodnoty θ při pravděpodobnosti 0,5 na hodnotu při pravděpodobnosti 0,5 + 0,5*c (Reeves 2002: 19, viz Graf 6).

Graf 6: ICC dichotomizované položky 34a (plat) dle 3PL modelu



Zdroj: vlastní výpočty v IRTPRO

1. 3. 3. Modely pro polytomní položky

Polytomní položky, tj. položky s více než dvěma kategoriemi odpovědi, jsou v sociologii velmi častým jevem, neboť mnoho sociologický otázek nejde rozřadit do prostých dichotomních odpovědí ano – ne, správně – špatně, které byly popsány v kapitole 0. Pro tyto polytomní odpovědi, které jsou často zastoupeny nejrůznějšími škálami, jsou nutné IRT modely, které reprezentují nelineární vztah mezi latentní proměnnou a pravděpodobností odpovědi v určité kategorii.

Kromě ordinálních škál, existují i modely pro nominální proměnné. Navíc existuje mnoho specifikací různých modelů, a nové každoročně přibývají (Embretson & Reise

2000: 95). Cílem této kapitoly je představit pouze základní modely, především ty, které lze provést pomocí dostupného software IRTPRO.

Model odstupňovaných odpovědí (Graded Response Model)

Autorkou modelu odstupňovaných odpovědí (GRM) je F. Samejima (1996) a jde v podstatě o rozšíření nám již známého 2PL modelu (viz kapitola 0). Tento model je možný použít pokud jsou odpovědi ordinální, například typicky Lickertova škála[8] či jiné typy posuzovacích škál.

Model odstupňovaných odpovědí je založen na logistické funkci, že odpověď na položku bude realizována v kategorii *k* či vyšší.

$$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ij})}}$$

V tomto modelu jsou zahrnuty právě dva parametry, stejně jako u 2PL modelu. Jediný rozdíl ve vzorečku oproti modelu 2PL představuje člen b_{ij}. Parametr *b*, obtížnost, se totiž liší mezi jednotlivými kategoriemi a dosahuje n-1 různých hodnot, tedy o hodnotu méně než je kategorií. Pro škálu se šesti kategoriemi, jako je v použitém příkladu, tedy odhadujeme parametrů *b* pět, přičemž platí $b_{k-1} < b_k < b_{k+1}$. Parametr *a*, diskriminace, je v tomto modelu shodný pro všechny odpovědi, což může být i pozorováno v grafu (viz. Graf 7), kde všechny křivky mají stejný sklon.

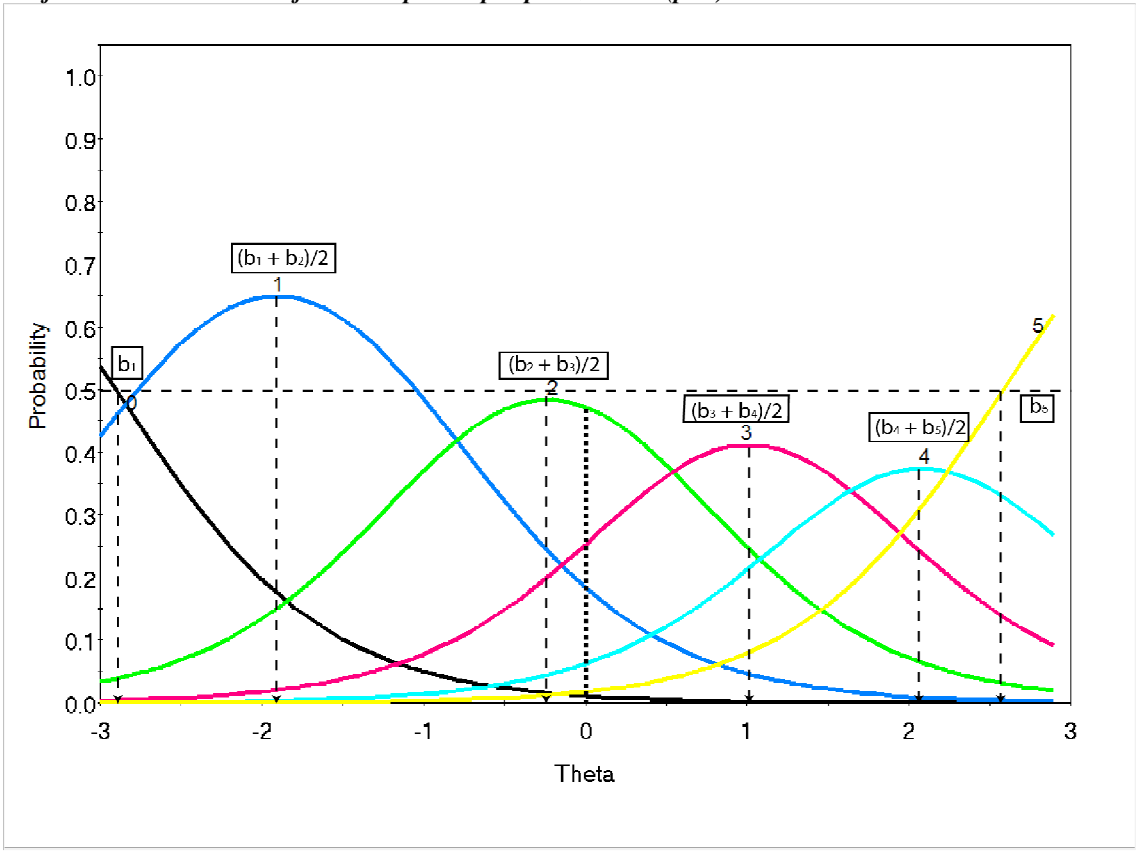
Pokud se podíváme na názorný příklad otázky z výzkumu (Vinopal 2011) – spokojenost s výší platu, mzdy, která je měřená na šestibodové škále (velmi spokojen až velmi nespokojen)[9] můžeme ze závislosti pravděpodobnosti určité odpovědi na latentní proměnné (viz Graf 7) vyčíst například, že pro střední hodnotu znaku (tj. $\theta = 0$, vyznačeno tečkovanou černou vertikálou) je pravděpodobnost krajních odpovědí velmi

[8] Lickertova škála je pětibodová stupnice, která měří úroveň ne/souhlasu.

[9] V dotazníku (Vinopal 2011) je použito kódování 1 – velmi spokojen až 6 - velmi nespokojen. Výstupy ze software IRTPRO vzhledem k potřebě odhadovat k – 1 parametrů jsou přečíslovány 0 až 5. Tedy funkce s legendou 0 (viz Graf 7) je funkce pro odpovědi velmi spokojen.

nízká – 1 % pro (0) „velmi spokojen“ a 2 % pro (5) „velmi nespokojen“. Vyšší pravděpodobnost (6 %) má odpověď (4) „nespokojen“, dále odpověď (1) „spokojen“ (18 %) a (3) „spíše nespokojen“ (25 %). Nejvyšší pravděpodobnost pro θ = 0 je, že respondent odpoví (2) „spíše spokojen“ (47 %).[10]

Graf 7: GRM model –IIC funkce odpovědí pro položku 34a (plat)[11]



Zdroj: vlastní výpočty v IRTPRO

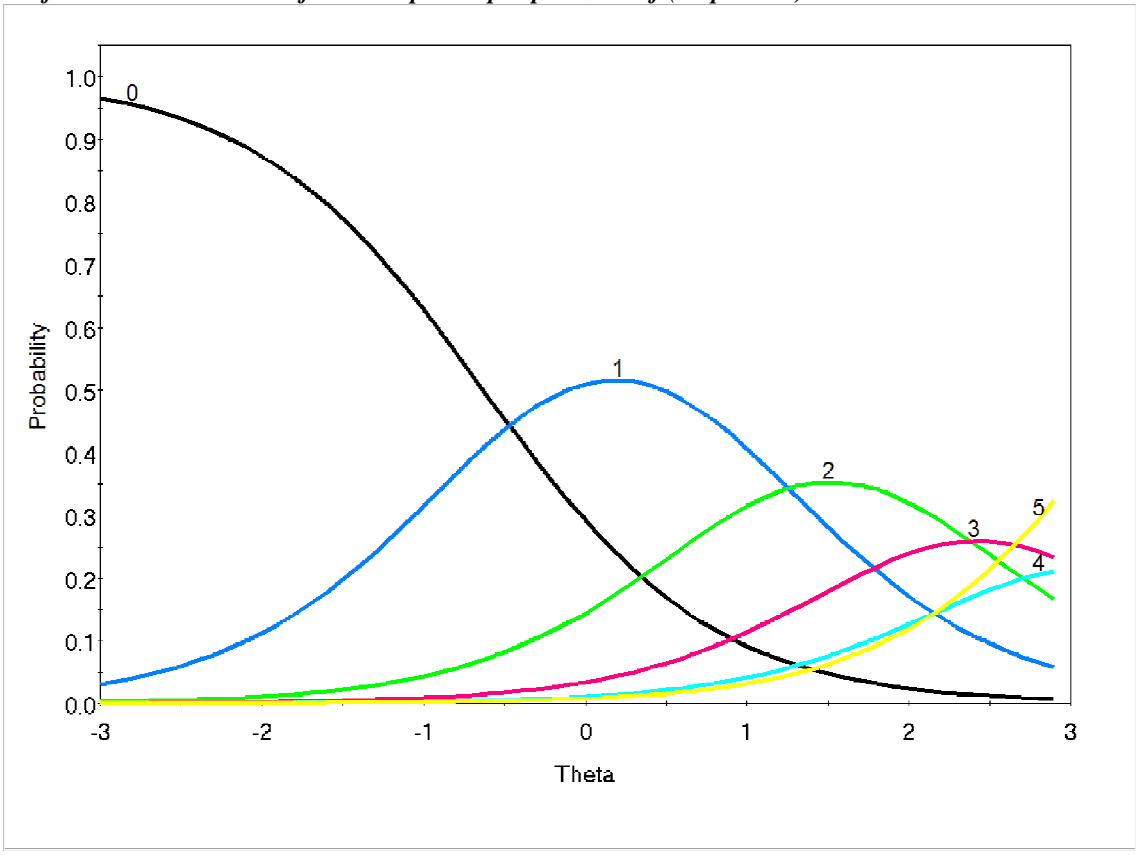
Odpověď (1) „spokojen“ je nepravděpodobnější při úrovni latentního znaku θ = (b1 + b2) / 2, tj. při úrovni θ = - 1, 9 a obdobně se tato pravděpodobnost dá vypočítat i pro další kategorie odpovědí (viz Graf 7). V ideálním případě je vždy pro

[10] Procenta u pravděpodobnosti odpovědí nedávají 100 %, ale 99 %. Jedná se o chybu vzniklou zaokrouhlením hodnot na celá čísla.

[11] V dotazníku měla tato i všechny ostatní položky ještě další dvě možné odpovědi nevím a netýká se. Tyto odpovědi byly pro názornost příkladu považovány jako tzv. missing values, nebyly tedy brány v potaz.

nějakou úroveň latentního rysu každá z šesti kategorií tou nejpravděpodobnější. Pokud tomu tak není (viz Graf 8), ukazuje to na nevhodně zvolené kategorie, protože respondenti mezi nimi nejsou schopni rozlišit. Konkrétní příklad v uvedeném grafu se týká otázky spokojenosti v práci – míra výskytu násilí a šikany. Odpověď (4) „nespokojen" je překryta ostatními odpověďmi. Graf také ukazuje kumulaci křivek vpravo, které tak dobře nerozlišují respondenty, kteří jsou nespokojeni s tímto aspektem práce. S nadsázkou (pokud vynechám „malý kousek" nejvyšší pravděpodobnosti křivky 3) by se dalo říci, že na tuto otázku respondenti odpovídají kategoriemi „spokojen" až „velmi spokojen". V opačném případě odpovědi nerozlišují a odpovědí rovnou nejextrémnější kategorií – „velmi nespokojen". Můžeme tedy shrnout, že původně navrhovaná šestibodová škála funguje spíše jako čtyřbodová.

Graf 8: GRM model – IIC funkce odpovědí pro položku 34f (bezpečnost)



Zdroj: vlastní výpočty v IRTPRO

Generalizovaný model pro stupňovaný kredit (Generalized Partial Credit Model)

Generalizovaný model pro stupňovaný kredit (Muraki 1997) je zobecněním modelu pro stupňovaný kredit (PCM – Partial Credit Model), jehož autorem je Masters (Masters & Wright 1997). Autoři vysvětlují PCM model na následujícím matematickém příkladu: $\sqrt{9/0{,}3-5}=?$

Aby byl tento příklad vyřešen, musí se po řadě:

(1) 9 vydělit 0,3
(2) odečíst 5
(3) odmocnit

Tedy nelze provést například krok tři, odmocňování, aniž by byly správně zodpovězeny kroky předchozí.

Model PCM má velmi restriktivní předpoklady. Jedná se totiž o rozšíření již známého 1PL modelu (viz kapitola 0), který má pouze jediný parametr, a to parametr *b*, obtížnost. Parametr *a,* rozlišovací schopnost jednotlivých položek, je považován za konstantní, což je v praxi jen velmi těžko splnitelné (Jelínek, Květoň & Vobořil 2011: 42). Z tohoto důvodu byl vyvinut generalizovaný model pro stupňovaný kredit (Muraki 1997), který je jedním ze základních výpočetních modelů i v programu IRTPRO, který je využíván v této diplomové práci.

Nominální modely (Nominal Category Model)

Nominální polytomní modely nepředpokládají uspořádané odpovědi, z čehož vyplývá, že špatné odpovědi jsou považovány za ekvivalentní, tedy za „stejně špatné“. Cílem analýzy pomocí nominálního modelu je právě určit, zda některé špatné odpovědi indikují rozdílné úrovně latentní proměnné (Drasgow & Hulin 1990: 596). Zda třeba ti s nižší úrovní matematických znalostí vybírají jako špatnou odpověď častěji variantu *a*, zatímco ti s vyšší úrovní, pokud nevyberou správnou odpověď, odpovídají spíše špatnou variantou *b*. Obecně můžeme říct, že nominální modely dávají do ruky výzkumníkům platný nástroj, jak určit diskriminační schopnost špatných (nežádoucích odpovědí). Dále tento model pomáhá určit různé odchylky v odpovídání respondentů či rozdíly mezi jednotlivými skupinami (Drasgow & Hulin 1990: 597). Například pro sociology

relevantnější příklad může být: Vybírají různé špatné odpovědi ženy a muži? Děti z města či z vesnice?

Jeden ze základních nominálních modelů je **Bockův model pro nominální kategorie** (Bock's Multivariate Logistic Model), který byl popsán v roce 1972 (Bock 1997: 24). Bockův model můžeme přiblížit jako zobecnění dvouparametrového logistického modelu s nominálními kategoriemi. Jako ilustrativní příklad využití v edukativních testech může být použito například hledání synonym určitého slova. Výzkumníci tímto dostávají nástroj, jak jednak i na základě špatných odpovědí usuzovat na úroveň latentní proměnné, tak nástroj, jak určit účinné „distraktory", nesprávné volby odpovědi pro danou otázku. U těchto modelů nabývá na důležitosti vizuální analýza grafického vyjádření jednotlivých odpovědí na položku, neb často mají různé tvary a je obtížné je popsat nějakou funkcí.

1. 3. 4. Další modely

Existuje celá řada dalších modelů vzniklých odvozováním či specifikací pro daný případ (například Miyazaki & Hoshino 2009). Mezi větší skupiny modelů patří také neparametrické modely či vícedimenzionální modely, jejichž podrobnější zpracování je již mimo rozsah diplomové práce i dostupného softwaru.

Neparametrické modely

V rámci teorie odpovědi na položku existují i neparametrické modely, pomocí kterých lze odhadnout informační a charakteristické křivky položky. Jedna z výhod použití neparametrických modelů je flexibilita ve tvarech křivek, data se nemusí „vtěsnat" do nějaké charakteristické křivky, jako má například dvouparametrový logistický model. Překonává tak problémy tzv. „lack-of-fit", kdy data nesedí do žádného parametrického modelu (Drasgow & Hulin 1990: 599). Zástupcem řady neparametrických modelů je například Levin's Multilinear Formula Scoring (MFS).

Vícedimenzionální modely

U vícedimenzionálních modelů se nejvýrazněji protíná teorie odpovědi na položku a klasická teorie testů, zastoupená faktorovou analýzou (Drasgov & Hulin 1990: 599), především tou nelineární (Embertson & Reise 2000: 83). Stejně jako u faktorové analýzy uvažujeme o dvou odlišných variantách, o explorativních (průzkumných) a konfirmativních (ověřovacích) modelech (Embertson & Reise 2010: 93).

Explorativní vícedimenzionální IRT model se používá se stejným cílem jako explorativní faktorová analýza, a to najít a interpretovat dimenze latentní proměnné. IRT má však tu výhodu, že nemá tolik přísné předpoklady (např. požadavek na normální rozdělení) a nedochází ke ztrátě informací na úrovni konkrétních položek u jednotlivých respondentů. Konfirmativní modely jsou pak používány pro návrhy propojení konkrétních položek s teoretickými předem danými latentními proměnnými. V současné době se vícedimenzionální modely používají především v úvodních fázích vývoje testu, při operacionalizaci (Embertson & Reise 2000: 83).

1. 4. Předpoklady IRT modelů

Stejně jako v klasické teorii testů, existují i pro teorii odpovědi na položku předpoklady, které musí být splněny pro správné používání a fungování modelů. Předpoklady DeMars (2010) shrnuje do třech základních požadavků – požadavek unidimenzionality, lokální nezávislosti a vhodnosti zvoleného modelu vzhledem k přirozené povaze dat. Dále bude diskutován i požadavek na velikost vzorku respondentů.

1. 4. 1. Předpoklad unidimenzionality[12]

Všechny možné modely, ať už dichotomní či polytomní, které jsou podrobněji popsány v kapitole 1. 3. mají podstatný společný rys. Předpokládáme, že odpověď respondenta je

[12] Existují i IRT modely, které pracují s vícedimenzionální latentní proměnnou, kde předpoklad unidimenzionality ztrácí své opodstatnění. Nicméně i u vícedimenzionálních modelů (viz zmínka v kapitole 1. 3. 4), musí být předem správně určen počet dimenzí.

ovlivněna pouze jedinou charakteristikou, tedy měřenou latentní proměnnou (v našem ilustrativním příkladu subjektivní spokojeností s pracovním životem). Ostatní proměnné, jakožto další individuální charakteristiky, vlivy prostředí aj. by dle předpokladu teorie odpovědi na položku neměly mít na odpověď respondenta žádný vliv. Z praxe však víme, že do reakcí respondenta vstupuje celá řada tzv. intervenujících proměnných, velkou roli také hraje například motivace, nervozita apod. (Jelínek, Květoň & Vobořil 2011: 50). Intervenující proměnné se ve výzkumu můžeme snažit minimalizovat, nicméně není možné je zcela potlačit. „Obecně se dá říci, že žádný soubor položek není striktně unidimenzionální" (Jelínek, Květoň & Vobořil 2011: 50). Modely IRT jsou poměrně robustní a tak stačí existence jednoho dominantního faktoru pro splnění předpokladu unidimenzionality.

Požadavek unidimenzionality lze testovat několika způsoby. (1) Nejznámějším je určování počtu dimenzí pomocí tzv. **eigenvalue**[13] či jejich zanesením do tzv. scree plot[14], což se běžně používá při faktorové analýze. Slabinou tohoto způsobu pro IRT je, že určení eigenvalues požaduje normalitu dat, což je silnější předpoklad, než je v IRT potřeba.

(2) Objektivnějším způsobem určování jednodimenzionality je tzv. **Stout's Test**, nazývaný také DIMTEST podle programu, ve kterém může být výpočet proveden (Stout 2005). Podstatou testu je rozdělení položek do dvou skupin, které jsou co nejvíce dimenzionálně odlišné, a následné otestování vytvořených skupin. V současné době málo rozšířenou, ale s velkou nadějí prozkoumávanou metodou testování dimenzionality, je (3) **analýza reziduí** (Demars 2010: 45).

---

[13] Eigenvalue používaná ve faktorové analýze je hodnota, která nám pomáhá určit počet faktorů. Její nejčastější referenční hodnotou je číslo 1. Faktor s nejvyšší eigenvalue vysvětluje nejvíce variability ze všech faktorů.

[14] Scree plot je čárový graf, který proti sobě dává eigenvalue a počet faktorů. Na základě vzniklé křivky a jejích zlomů se posuzuje počet faktorů v datech.

Lze shrnout, že se jedná o vyvíjející se oblast zkoumání a zatím není stanovený „obecně uznávaný“ postup, jak rozhodnout, zda je splněn předpoklad unidimenzionality. Analyzovaná data nebudou nikdy striktně unidimenzionální a i model je vždy jen nedokonalým přiblížením reálných dat. Proto je tak důležité vyvinout statistiky, které pomohou určit, zda model splňuje předpoklad unidimenzionality, aby odhady parametrů byly konzistentní a reliabilní. Nicméně v současnosti dle Embertson & Reise (2000) výzkumníci spíše vědí, jaké testy nejsou vhodné, jaké statistiky nepoužívat, než aby dokázali formulovat jednoznačná doporučení, jak postupovat. Například je známo, že velikost indexů interní konzistence (např. alfa koeficient) je pro určování dimenzionality irelevantní (Embertson & Reise 2000: 231).

1. 4. 2. Předpoklad lokální nezávislosti

Odpovědi na jednu položku by měly být nezávislé na odpovědi na jinou položku, pokud kontrolujeme θ. Jinými slovy korelace mezi dvěma položkami v dotazníku by měla být způsobena pouze latentní proměnnou θ, nikoliv jinou společnou proměnnou, která není zahrnuta v modelu (DeMars 2010: 37). Předpoklad lokální nezávislosti je splněn v případě, že se v datech systematicky neobjevují jiné kovariance, pokud kontrolujeme latentní proměnnou θ.

Empiricky bylo zjištěno, že následky porušení předpokladu lokální nezávislosti se projevují v odhadu parametrů a informačním přínosu testu, které jsou pak větší, než by měly být (Embertson & Reise 2000: 232). Doporučený přístup, jak řešit problematiku lokální nezávislosti, je jí předcházet už při přípravě výzkumu. Zjednodušeně a z praktického pohledu lze tvrdit, že pokud je splněn předpoklad unidimenzionality, tak je pravděpodobně splněn i předpoklad lokální nezávislosti (Orlando 2004). Pozor je potřeba dát pouze v případě, kdy předchozí položky poskytnou vhled, který usnadní či jinak ovlivní řešení dalších položek.

1. 4. 3. Určení vhodného modelu

V rámci teorie odpovědi na položku existují desítky modelů, které se mohou i alternovat. Na základě teoretické úvahy se vybere nejvhodnější model či modely, které

je pak nezbytné otestovat, zda předpokládané odpovědi na položku korespondují s těmi empiricky zjištěnými.

Ještě než přistoupíme ke konkrétním technikám zjišťování vhodnosti modelu, je nezbytné se zabývat obecně modelem samým. Každý z modelů vytvořených na základě teorie odpovědi na položku je popsán minimálně jedním parametrem (více viz kapitola 1. 2. Existuje několik způsobů, jak se parametry z dat odhadují, a to jak parametry osob, tak položek, neboť většinou není známé ani jedno (Jelínek, Květoň & Vobořil 2011: 61). Nejpoužívanější metodou, jak z pozorovaných dat (*n* položek určitého testu a *N* odpovědí na test) odhadnout parametry modelu, je metoda maximální věrohodnosti (Embertson & Reise 2000: 187). Mezi další metody odhadu patří např. známá metoda nejmenších čtverců používaná v lineární regresi či Bayesova metoda.

Podstata metody maximální věrohodnosti (Maximum Likelihood) spočívá ve výběru nejpravděpodobnější odpovědi pro danou úroveň latentní proměnné. Tedy konkrétně pro naše data, že respondent s určitou úrovní subjektivní spokojeností s pracovním životem vybere některou z odpovědí na otázku velmi pravděpodobně, jinou méně pravděpodobně a některou je téměř nemožné, aby vybral.

Existuje několik způsobů odhadu maximální věrohodnosti. (1) Spojený odhad maximální věrohodnosti (Joint Maximum Likelihood), marginální (Marginal Maximum Likelihood) a podmíněný (Conditional Maximum Likelihood). Konkrétní matematické vyjádření, které není žádoucí součástí této diplomové práce, je možné běžně dohledat v literatuře zabývající se IRT (např. Embertson & Reise 2000: 8. kapitola) či statistickou analýzou (např. Hendl 2009).

Nyní lze přikročit k určování vhodnosti modelu pro daný účel a data. Při výběru modelu samozřejmě v první řadě závisí na tom, zda jde o položku dichotomní či polytomní. Mezi dichotomními modely je nejběžnější tříparametrový logistický model (3PL) pro tzv. „multiple choice“ odpovědi. Pro položky měřící postoje s právě dvěma odpověďmi zase dvouparametrový (2PL). Dále, kromě počtu odpovědí, je důležitým kritériem, jak

odpovídá model datům. Nejen pro polytomní modely se doporučuje použít jednodušší precizněji odhadnutý model (DeMars 2010: 30).

Často se ale stane, že i přes prvotní filtr na základě vlastností položek, je k dispozici více alternativních modelů, ze kterých je třeba vybrat ten nejadekvátnější. Dle Jelínek, Květoň & Vobořil (2011) se vhodnost modelu dá testovat na několika úrovních. Na úrovni (1) položek (item fit), (2) specifickými indexy na úrovni osob (person fit) nebo (3) modelu jako celku (model-fit), a to jak graficky, formálně či kombinací obou přístupů.

Nicméně není zaveden žádný set indikátorů, který by byl uznáván pro měření vhodnosti vybraného modelu (Reeve & Fayers 2005: 63). Velkou roli zde tedy hraje také odbornost a úsudek výzkumníka. Přesto existuje jedno nepsané pravidlo, a to, že přirozená povaha dat musí co nejlépe odpovídat zvolenému modelu při zachování co největší jednoduchosti modelu (Demars 2010: 37).

(1) Vhodnost modelu na úrovni položek

Při aplikaci teorie odpovědi na položku na výzkum není nutné používat stejný model pro všechny položky dotazníku. Například test se může skládat z kombinace dichotomních a polytomních položek, některé můžou být reprezentovány tříparametrovým logistickým modelem, jiné nominálním modelem. Proto je důležitější, jak už i z názvu teorie odpovědi na položku vyplývá, najít vhodný model pro jednotlivé položky než jednotný model pro celé dotazníkové šetření.

Existují v podstatě dva přístupy hodnocení, jak dobře IRT model vysvětluje či předpovídá odpovědi na konkrétní položku (Embertson & Reise 2000: 234). Jeden z přístupů je více heuristický či grafický, kde nejsou aplikovány žádné statistické testy. Druhý přístup je více formální.

(1) Jedním z heuristických přístupů je porovnání, jak dobře odpovídá charakteristická křivka položky pozorovaným reálným datům. Po odhadu parametrů se určí hodnoty pro

jednotlivé respondenty a seřadí se vzestupně a rozdělí nejčastěji do deseti stejně početných skupin. Následně se střední hodnoty těchto vypočítaných skupin porovnají s reálnými. Odchylky a špatná shoda s modelem dle Embertson & Reise (2000: 234) nejčastěji zapříčiňuje (a) nesplnění předpokladů modelu, jako je například unidimenzionalita, (b) nedostatečný počet odhadnutých parametrů (například aplikování jednoparametrového logistického modelu na místě, kde má své opodstatnění dvouparametrový). Mezi další důvody špatné shody patří, že (c) charakteristická funkce položky není monotónní či (d) je mezi respondenty nějaká značně odlišná subpopulace či skupina, pro kterou nástroj neměří standardně, popřípadě (e) položka byla špatně formulována a odpovědi na ní nejsou tak konzistentní.

(2) Kromě těchto vizuálních, grafických technik je snaha i formalizovat srovnávání empirické a modelové funkce položky pomocí statistik pro signifikanci reziduí. Jedním z takových pokusů je například, jak uvádějí Embertson & Reise (2000: 235), Bock chi-square, či jiné chi-square přímo implementované do statistických programů využívaných v teorii odpovědi na položku[15], popřípadě je snaha i o využití standardizovaných reziduí.

(2) Vhodnost modelu na úrovni osob (respondentů)

Statistiky sloužící pro hodnocení modelu na úrovni osob zkoumají validitu modelu na individuální úrovni respondenta. Pro výzkum tzv. „person-fit" existují různé statistiky (například z_3, F_2), některé z nich jsou použitelné pouze pro konkrétní modely, které jsou však vesměs založeny na konzistenci odpovědí respondentů s navrhovaným modelem (Embertson & Reise 2000: 235). Nicméně posuzování vhodnosti modelu na úrovni osob není běžný rutinní krok v IRT analýzách (DeMars 2010: 57).

[15] Statistických programů, které jsou používané v rámci teorie odpovědi na položku, je velká řada. Nicméně ve velkém množství případů se jedná o komerční programy, které neposkytují akademické licence. Také jsou programy často vyvinuty právě výzkumníky, kteří si do nich implementují své statistické testy, např. BILOG 3 (Mislevy & Bock, 1990) či BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996)

Zkoumání konzistence odpovídání respondentů lze však také využít pro diskuzi validity dotazníku. Systematické odchylky odpovědí od modelu mohou být způsobeny právě tím, že měří něco jiného. Například jak uvádí Jelínek, Květoň & Vobořil (2011: 83) „nespavost je v rámci škály neuroticismu považována za výrazný symptom tohoto rysu, ačkoliv může být příznakem nesouvisejících potíží."

(3) Vhodnost modelu jako celku

Model jako celek se také zdaleka neposuzuje v IRT analýzách tak často a běžně jako na úrovni jednotlivých položek (DeMars 2010: 57). Obecně jde o statistiky, které jsou založeny na chí kvadrátu, který porovnává rozložení skórů jednotlivých testů, či využívá statistiky maximální věrohodnosti, pokud je použita pro odhad parametrů (DeMars 2010). Nicméně se jedná o velmi komplexní problematiku, která je předmětem aktuálního výzkumu a nezbývá na ní již v této diplomové práci prostor.

1. 4. 4. Počet respondentů

Odpověď na otázku, kolik respondentů je potřeba pro vytvoření kvalitních modelů v rámci IRT, není jednoduchá ani jednoznačná. Velikost vzorku respondentů je ovlivněna (1) výběrem modelu. Obecně platí, že čím složitější model s více parametry, tím větší je třeba vzorek. Pro jednoduché Rasch modely někdy stačí i pouhých 50 respondentů. Proto je tento model také oblíben u výzkumníků v lékařství, neboť je často nereálné sehnat více pacientů (Reeve & Fayers 2005: 71). Naopak polytomní modely vyžadují logicky větší vzorek. Například u modelu odstupňovaných odpovědí (GRM) bylo ukázáno, že stačí 250 respondentů, ale pro přesné odhady parametrů se doporučuje velikost okolo 500 respondentů (Embertson & Reise 2000: 123).

Velikost vzorku také může záviset na (2) účelu výzkumu. Pro zhodnocení vlastností dotazníku není třeba příliš velký vzorek, ale zase je důležité, aby byl heterogenní a co nejlépe pokrýval celé kontinuum. Ale pokud je účel kalibrovat položky do banky pro počítačové adaptivní testování (více v kap. 2. 3.), tak jsou třeba vzorky ještě větší než 500 respondentů (Embertson & Reise 2000: 123).

Dalším důležitým kritériem je (3) potřebná velikost standardizované chyby, která je pro účel výzkumu ještě akceptovatelná. A to především v extrémních kategoriích, které nejsou tolik frekventované. Velikost vzorku také závisí na kvalitě splnění požadavků modelu popsané výše v této kapitole - unidimenzionality, lokální nezávislosti a vztahu mezi měřenými proměnnými a latentním konstruktem (viz kapitola 1. 4. S rostoucím počtem možných odpovědí v baterii otázek, roste počet parametrů, které je třeba odhadovat, což opět zvyšuje požadavky na velikost vzorku (Embertson & Reise 2000: 123).

1. 5. Validace výsledků

Otázka validity zkoumá, zda daný konstrukt měří správně latentní proměnnou, laicky řečeno, zda měří to, co měřit má. Konkrétně pro naši případovou studii validita znamená, zda set osmnácti otázek opravdu měří subjektivní spokojenost s pracovním životem (Vinopal 2011). Protože, pokud by zde nebyla souvislost, vypovídající hodnota jakkoliv přesných a dokonalých modelů by byla přinejmenším pochybná.

Abychom zjistili validitu nástroje a zjištěných výsledků, je potřeba provést nějakou studii validity. Existuje několik nástrojů a postupů k měření validity. (1) Zjevná validita je spíše intuitivní metoda, (2) obsahová validita se odkazuje na existující teorie, literaturu či experty (Kreidl 2005). Ve vzdělávací oblasti se například může monitorovat (3) kriteriální validita, zda výsledek u přijímacích zkoušek pomáhá předpovídat budoucí úspěch při studiích (Johnson 2004: 24). Posledním přístupem je (4) konstruktová validita, která je založena na zjištění takového vztahu mezi proměnnými, který je i doložitelný teorií (Kreidl 2005). Validita nástroje pro výzkum subjektivní kvality pracovního života byla průběžně ověřována při jeho konstrukci dalšími otázkami v dotazníku (úvaha o změně zaměstnání, celková spokojenost se zaměstnáním – více viz Vinopal 2011).

Ačkoliv je IRT jedním z nejpřínosnějších metodologických přístupů posledních let, tak nebylo věnováno příliš pozornosti jeho vlivu na zvýšení kvality a validity výzkumů

(Linn 1989), což platí dodnes. Nicméně pro měření právě konstruktové validity může mít teorie odpovědi na položku významný přínos, a to především při (1) vytváření škál, (2) výběru položek do dotazníku či zkoumání (3) vlivu instrukcí na odpověď respondenta (Linn 1989). IRT dále přispívá k měření validity vytvořenými modely, a to především těmi, které „nesedí". Pomáhá odhalovat možné problémy s validitou, jak bylo například zmíněno u měření míry neuroticismu a nespavosti v minulé kapitole (1. 4. 3) či různé systematické chyby mezi skupinami (sklon k depresím a pláč v kapitole 2. 2.

Teorie odpovědi na položku tak může výrazně přispět k vylepšení výzkumného nástroje, který je vytvořen na základě nenahraditelného teoretického konceptu, zkušeností výzkumníka a precizní formulace otázek. Studie použití IRT a jejího vlivu na validitu výzkumu by byla vhodným a zajisté dostatečně ambiciózním rozšířením této diplomové práce, neboť jak říká Linn (1989) je v této oblasti stále „více otázek než odpovědí".

1. 6. Srovnání IRT a CTT

Klasická teorie testů i teorie odpovědi na položku mají mnoho společného i odlišného, u každé můžeme najít některé výhody i nevýhody. Na tomto místě je třeba zdůraznit, že tyto teorie, ač by se tak na první pohled mohlo zdát, nestojí proti sobě, používání jedné nevylučuje použití druhé. I přes následné popsané výhody teorie odpovědi na položku by se rozhodně nemělo sklouznout k přílišnému nadšení z matematických modelů. Ty jakkoliv propracované stále slouží spíše jako další pomůcka pro vytváření kvalitních dotazníků a rozhodně nenahrazují klasickou teorii testů a pečlivou práci při formulování otázek.

Hlavní rozdíl mezi CTT (klasickou teorií testů) a ICT (teorií odpovědi na položku) leží právě v postavení položky testu[16]. V klasické teorii testů je položka chápána jako

[16] Test v této souvislosti je chápán jako soubor otázek měřící jeden konstrukt.

neoddělitelná součást konkrétního testu. Jinými slovy, bez kontextu celého testu nejsou v CTT nástroje, kterými by bylo možné zjistit, zda položka měří to, co skutečně měřit má. Zatímco v IRT se pohlíží na položku individuálně a sledují se její vlastnosti (Urbánek & Šimeček 2001: 428). Z této povahy samostatné položky v IRT vyplývá několik praktických vlastností a aplikací (viz kapitola 2. 1.

1. 6. 1. Reliabilta

Jedním ze základních rozdílů mezi IRT a CTT je přesnost měření některých statistik, například reliability či směrodatné odchylky.

(1) Směrodatná odchylka (SEM) měří očekávané rozdíly, jak moc se námi získané údaje ze statistického vzorku liší od základního souboru díky nedokonalosti měřícího instrumentu. Je logicky žádoucí ji minimalizovat. Klasická teorie testů (CTT) předpokládá, že standardní chyba je konstantní pro všechny úrovně θ. Oproti tomu v teorii odpovědi na položku (IRT) se tato míra odhaduje odděleně pro každou úroveň latentní proměnné. Typicky je preciznost měření nejlepší uprostřed kontinua a nejmenší na obou koncích, kde test nerozlišuje kvalitně mezi respondenty (Reeve 2002: 32). Proto je také méně vhodné používat v tomto případě klasickou teorii testů, neb v realitě distribuce směrodatné odchylky není stejná pro všechny úrovně latentní proměnné, jak implicitně CTT předpokládá.

(2) Reabilita ve smyslu interní konzistence testu je často popisována statistikou Cronbachovo alfa, která je založená na předpokladu, že by všechny položky měřící konstrukt měly mít mezi sebou vysoké korelace (Hendl 2009). Analogií reliability je v IRT informační křivka položky (viz Graf 4 nebo Graf 5), neboť s rostoucí informací klesá chyba měření. Rozdíl spočívá v tom, že v klasické teorii testů se reliabilita nehodnotí na úrovni jedné položky, ale pro celou skupinu skórů (DeMars 2010: 82).

Teorie odpovědi na položku, konkrétně v modelech, které nevycházejí z Raschova modelu[17], bere navíc v potaz obtížnost položky a její schopnost diskriminovat (Reeve & Fayers 2005: 61). Položky, které více přispívají k celkové informační křivce testu (podrobněji viz Graf 9 a kapitola 2. 1. tedy ty, které jsou více diskriminující, či mají větší reliabilitu, mají v konečném modelu implicitně větší váhu. Výsledek testu pak může mít vyšší reliabilitu, než kdyby byly položky analyzovány klasickou teorií testů bez přidání vah (DeMars 2010: 3). Problematika vážení položek v IRT a i CTT, oblast explicitních i implicitních vah (parametrů) je předmětem současného výzkumu, pro studium lze doporučit například Stucky 2009.

S vážením jednotlivých položek souvisí i další oblast, a to (3) různé druhy položek v rámci jednoho testu, dichotomní položky, škály, otevřené odpovědí. V klasické teorii testů mají různé formáty položek rozdílný vliv na celkový výsledek testu. Položky s velkým počtem odpovědí tak často nejvíce ovlivňují výsledek celého testu (Reeve 2002: 33). V kontrastu s CTT existují v rámci IRT různé modely jak pro dichotomní tak polytomní, ordinální i nominální položky. V celkovém IRT modelu mohou být odlišné modely položek, což řeší problematikou smíšených formátů otázek a jejich vah podstatně elegantněji.

V klasické teorii testů je samozřejmé, že (4) delší testy mají vyšší reliabilitu (Embertson & Reise 2000: 18). V rámci IRT však vhodným výběrem položek, které pokrývají celé kontinuum, a mají nejvyšší rozlišovací účinnost, lze realitu CTT zvrátit. Jedná se však především o situaci, kdy je použito adaptivní testování (více viz kapitola 2. 4. které umožňuje respondentovi zodpovídat jen takové otázky, jež jsou pro měřenou úroveň latentního rysu nejrelevantnější.

---

[17] Reeve a Fayers (2005) do rodiny Raschových modelů zařazují jednoparametrový logistický model, model pro stupňovaný kredit a model odstupňovaných odpovědí. Naopak modely, které nevycházejí z Rasche, jsou dvou a tříparametrový logistický model, nominální model a zobecněný model pro stupňovaný kredit.

1. 6. 2. Nezávislost parametrů položek a osob

Jedním ze základních principů IRT je nezávislost parametrů položek a osob. To znamená, že parametry položek jsou neměnné napříč skupinami, populacemi osob. Parametry osob jsou zase při vhodně zvoleném modelu nezávislé na použitých položkách (Jelínek, Květoň & Vobořil 2011: 99). Z principu nezávislosti parametrů položek a osob vyplývá několik praktických výhod usnadňujících a zpřesňujících výzkum.

Jednou z užitečných výhod je, že testy vypracované na základě IRT jsou (5) relativně nezávislé na skupinách respondentů, na testované populaci (DeMars 2010: 7), zatímco v klasické teorii testů se obtížnost položky změní, pokud test podstoupí další respondent s vyšší či nižší hodnotou θ. Tato výhoda by však neměla být přeceňována, protože pokud položky nejsou příliš extrémní či skupiny příliš odlišné, funguje dobře i CTT. Díky této nezávislosti již není nutné interpretovat výzkum pouze v kontextu populace, na které byly položky kalibrovány. Jako příklad lze uvést parametr *b,* obtížnost položky a jeho ekvivalent v klasické teorii testů, podíl klíčových odpovědí na položku, takzvaná míra jednoduchosti[18]. Pokud test budou vyplňovat jedinci s nadprůměrnou hodnotou latentní proměnné, míra jednoduchosti se zvýší. V IRT lze nezkreslené odhady vlastností položek získat i z nereprezentativních výběrů. Tato vlastnost se nazývá invariance (Kujal 2008).

Jako výhoda je oceňováno také (6) vyjádření parametrů položek a osob na stejné škále, a to na škále, na které je měřená latentní proměnná (Jelínek, Květoň & Vobořil 2011: 99). Takovéto vyjádření rozšiřuje možnosti interpretace výzkumu. Navíc přenesení těchto informací do grafické podoby (viz Graf 10) poskytuje výborné možnosti pro prezentaci výsledků, rozhodování a je srozumitelné i osobám, které neznají IRT.

[18] Míra jednoduchosti položky je pracovní název pro podíl správných, resp. žádoucích odpovědí na položku a celkového počtu odpovědí na danou položku.

1. 6. 3. Předpoklady

Kujal (2008) považuje podmínky pro použití postupů CTT jako „slabé“, tj. testovými daty snadno splněny a IRT jako „silné“, testová data je splní obtížněji. Jako příklad uvádí požadovanou unidimenzionality dat, pokud chceme použít jednodušší, unidimenzionální modely.

Na druhou stranu mnohá úskalí při používání klasické teorii testů představuje požadavek normality. Ten v rámci IRT není formulován, data nemusí mít normální rozdělení (7). Pro odhady parametrů z položek není podle teorie odpovědi na položku oproti klasické teorii potřeba reprezentativní výběr z cílové populace. V IRT může být i (8) nereprezentativní, musí však splňovat požadavek, že respondenti dostatečně pokryjí celou škálu θ, aby mohly být parametry modelu co nejpřesněji odhadnuty (DeMars 2010: 32).

Nejen rozdíly v požadavcích, ale ve zjednodušené podobě jsou v následující tabulce stručně a přehledně shrnuty některé rozdíly mezi klasickou teorií testů a teorií odpovědi na položku (viz Tab. 2).

Tab. 2: Srovnání IRT a CTT

	Klasická teorie testů (CTT)	**Teorie odpovědi na položku (IRT)**
	Reliabilita	
(1)	Směrodatná odchylka je konstantní pro různou úroveň θ.	Směrodatná odchylka se liší pro různou úroveň θ
(2)	Cronbachovo alfa je konstantní pro různou úroveň θ.	Informační křivka položky se liší pro různou úroveň θ
(3)	Smíšené formáty položek mají nevyvážený vliv na výsledek testu.	IRT snáze řeší problematiku smíšeného formátu položek.
(4)	Delší dotazníky zvyšují reliabilitu.	Kratší dotazník může mít větší reliabilitu než delší.
	Vlastnosti položek	
(5)	Položky a její vlastnosti jsou závislé na kontextu testu (např. podíl odpovědí na položku), nelze je tedy zobecnit	Položky a její vlastnosti jsou nezávislé na kontextu testu (např. parametr *b*), lze je tedy zobecnit.
(6)		Vyjádření parametrů položek a osob na stejné škále.
	Předpoklady	
(7)	Předpoklad normality*	Předpoklad unidimenzionality*
(8)	Reprezentativní výběr	Nereprezentativní výběr

Zdroj: Vlastní zpracování na základě kapitoly 1. 6. A literatury v ní použité.
* předpoklad neplatí pro všechny modely / statistiky

2. Aplikace IRT v sociologii a její srovnání s klasickou teorií testů

Teorie odpovědi na položku je alespoň v české sociologii zatím opomíjenou teorií, spíše se prosadila v oblasti vzdělávání a psychologie při sestavování testů. Cílem této kapitoly není přinést vyčerpávající seznam aplikací IRT, ale spíše ukázat možnosti využití této teorie i v sociologickém kontextu a posoudit její přínosy.

Mezi příkladné důležité otázky, které řeší sociologové při tvorbě kvalitních dotazníků, patří například: Pokrývají položky a kategorie škály dobře celé kontinuum měřené latentní proměnné θ? Nebo jsou některé kategorie nadbytečné, překrývají se a respondenti mezi nimi nejsou schopni rozlišit? Jakou funkci má v dotazníku každá položka, jaká je její rozlišovací schopnost? Jak přispívá k celkové informaci získané testem? Je dotazník adresován správně? Není pro respondenty příliš obtížný nebo jednoduchý? Jinými slovy odpovídá rozložení obtížnosti položek v dotazníku rozložení latentní proměnné θ v populaci (inspirováno Demars 2010)? Teorie odpovědi na položku umožňuje takovéto otázky na základě modelů uspokojivě zodpovídat a přispívat tak k vyšší kvalitě sociologického výzkumu.

2. 1. Analýza položek – vytváření krátkých a reliabilních výzkumných nástrojů

Teorie odpovědi na položku poskytuje metodologii, jak zkoumat odděleně charakteristiky každé položky výzkumného šetření a určit (1) jaké úrovně latentní

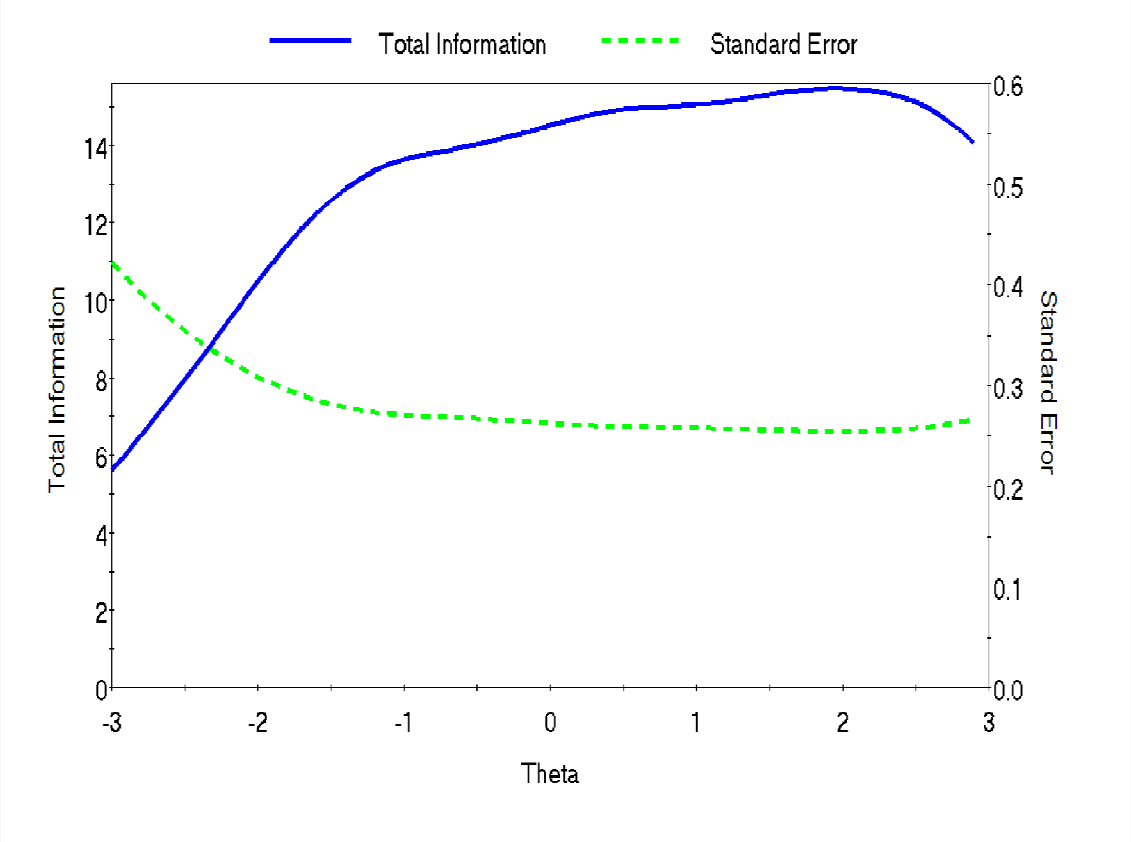
proměnné měří a do jaké míry je položka (2) přínosná pro výzkumný nástroj. Pro provádění IRT analýz hrají podstatnou úlohu různé grafické výstupy modelů.

Právě druhý zmíněný pohled diskriminace položky reprezentovaný parametrem *a*, pomůže odhalit položky, které rozlišují mezi respondenty dobře a které špatně. Informační křivka položky (ukázaná na Graf 4 a také Graf 5) zase ukazuje, jak velkou informaci položka přináší pro různé hladiny měřené latentní proměnné, v našem případě subjektivní spokojenosti s pracovním životem. Nedostatečně měřící položky mohou být vyloučeny při nezměnění, ba zlepšení reliability a tím lze celý výzkumný nástroj zkrátit, což je žádoucí jak pro výzkumníky, tak pro respondenty (Reeve 2002: 34)

Pokud se jedná o baterii otázek se stejnou škálou odpovědí, na které lze použít stejný model, můžeme získat a porovnávat i informační křivku celého souboru položek, testu jako celku. Výsledný grafický výstup, informační přínos testu (Total Information Curve – TIC) nám pomůže určit, kde nástroj měří nejlépe. Například z celkové informační křivky baterie (viz Graf 9) je patrné, že test obsahuje většinu položek, jež měří lépe respondenty, kteří nejsou tolik spokojeni[19] se svým pracovním životem, což se také projevuje na celkovém informačním přínosu testu (vrchol modré křivky – „total information“). Chyba měření (zelená křivka – viz Graf 9) je nejvyšší právě pro takové respondenty, kteří jsou subjektivně nejvíce spokojeni se svým pracovním životem a pro které je takto postavený test nejméně adekvátní.

[19] Odpovědní škála je šestibodová (1 – velmi spokojen, 6 – velmi nespokojen), proto na vodorovné ose x (Theta) na Graf 9 jde latentní proměnná zleva doprava od velmi spokojen až po velmi nespokojen.

Graf 9: GRM model – Celková informační křivka (TIC) a chyba měření baterie



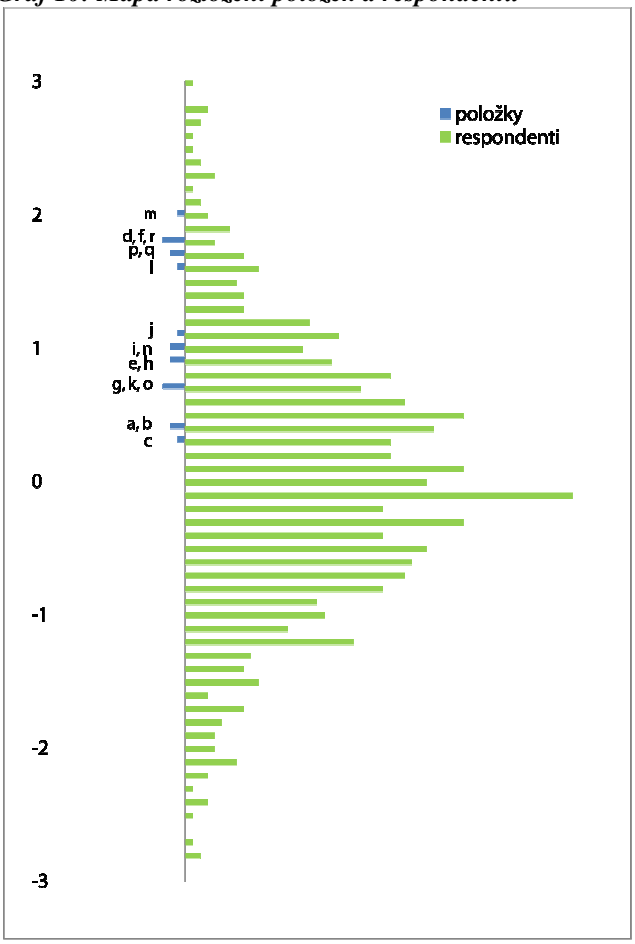
Zdroj: vlastní výpočty v IRTPRO

Za další z výhod IRT se považuje možnost zobrazit rozložení respondentů i obtížnosti položek na jedné ose. Jeden z nejběžnějších a snadno interpretovatelných nástrojů umožňující žádoucí zobrazení je tzv. položková mapa. Toto znázornění se v některé literatuře nazývá „Wright map" (DeMars 2010: 76), běžněji jako „item map". Graf 10 jednu takovou mapu ukazuje. Na levé straně je vyznačená poloha položek dle jejich „obtížnosti" (parametr *b*), na pravé straně grafu je ukázána distribuce respondentů, v našem případě dle subjektivní (ne) spokojenosti s pracovním životem[20]. Každá část úsečky[21] představuje jednu položku/respondenta.

[20] Subjektivní spokojenost s pracovním životem je v grafu dle kódování trochu neintuitivně vyznačena nižšími čísly na vertikální ose. Čili čím vyšší číslo, tím vyšší subjektivní nespokojenost s prací.

[21] Vhodnějším grafickým vyjádřením by byl bod, křížek apod., nicméně používaný software IRTPRO mapu položek negeneruje, proto byla dělána ručně.

Graf 10: Mapa rozložení položek a respondentů



Zdroj: vlastní výpočty v IRTPRO

Umístění položky na kontinuu slouží k určení, zda je výzkum správně zacílen, zda nástroj dobře rozlišuje tam, kde potřebujeme. Na mapě (Graf 10) vidíme, že průměrná „obtížnost" položek se realizuje především v horní polovině grafu, tedy že měří především respondenty, kteří jsou spíše nespokojeni s prací. Pokud předložená baterie má sloužit jako univerzální měřící prostředek, bylo by vhodné přidat či přeformulovat některé z otázek, aby lépe měřily i vyšší subjektivní spokojenost[22].

[22] Toto doporučení je pouze hypotetické na základě grafu. Software IRTPRO nepodporuje tento typ grafu a proto byl pro názornost alespoň vytvořen ručně z dichotomizovaných položek.

Další informací, kterou na mapě (viz Graf 10) zjistíme, je, že položky *d, f, p, q* a *r* měří respondenty s podobnou mírou latentní proměnné, poskytují tak redundantní informace. Doporučení na základě zobrazené mapy by bylo odstranit některé ze zmíněných položek a nahradit je těmi, které měří i v druhé části kontinua, tedy respondenty nadprůměrně spokojené se svým pracovním životem.

Mapu položek lze také využít pro diskutování validity, a to především u výzkumných šetření, kde mají položky přesné pořadí, typicky kumulativní Guttmanovy škály[23], jejichž pořadí může být v rámci teorie odpovědi na položku potvrzeno a dále analyzováno.

Analýza položek se samozřejmě provádí i v rámci klasické teorie testů, například jejich obtížnost jako „podíl správných odpovědí na položku“ či bodově-biseriální koeficient (více viz Tab. 1), nicméně právě s doplněním o modely IRT dostávají výzkumníci silný nástroj pro vytváření optimálních dotazníků pro tvorbu krátkých, reliabilních škál, které jsou vhodné pro studovanou populaci (Reeve & Fayers 2005: 57). Je umožněno také určit optimální počet odpovědí pro jednotlivé položky.

2. 1. 1. Porovnávání reliability výzkumu při různých módech sběru dat

S možností analyzovat položky odděleně se pojí i další ze sociologicky relevantních využití IRT, a to v oblasti sběru dat, tzv. mixed modes. IRT pomáhá určit, zda se liší pravděpodobnosti odpovědi na otázky respondentů se stejnou úrovní latentní proměnné při použití odlišných technik dotazování (např. klasické pomocí „tužky a papíru“ - PAPI, telefonu - CATI či počítače - CAPI)[24] a zda jsou takové výzkumy srovnatelné.

[23] Známy příklad Guttmannovy škály je např. škála sociální distance (Bogardus 1933), která se ptá do jaké míry, by respondent přijal člena jiné sociální skupiny – od svatby až po vyhoštění ze země.

[24] PAPI (paper and pencil interviewing) – osobní dotazování, při kterém se odpovědi respondentů zaznamenávají do papírového dotazníku; CATI (computer assisted telephone interviewing) – telefonické

S rozvojem moderních technologií se u mnoha výzkumů začínají využívat počítače (Xu – Iran-Nejad – Thoma 2007: 10). Technicky není složité navrhnout a převést výzkum do elektronické podoby, otázkou však zůstává, zda změna sběru dat nějak ovlivní reakce respondentů, což nabývá na důležitosti především u opakujících se výzkumů (např. některé ze šetření Naše společnost[25]).

Ferrando & Lorenza-Seva (2005) shrnuli rozdíly mezi klasickou „papírovou" verzí sběru dat a moderní počítačovou verzí sběru dat, často za pomocí internetu, do třech základních oblastí. (1) U anonymních internetových výzkumů je často problematická povaha vzorku respondentů. Ale i pokud je vzorek shodný, stále je nutné brát v potaz, že (2) se liší způsob předání informací. Vstupují sem faktory, jako je například vztah k počítačům, změna kvality čtení otázky apod. (3) Třetí oblastí jsou pak podmínky, za kterých je dotazník respondentem vyplňován – okolí, uživatelské rozhraní dotazníku a další.

Současný výzkum porovnávání shody výsledků při různém sběru dat nahlíží na problematiku jak z pohledu klasické teorie testů, tak teorie odpovědi na položku. Klasická teorie testů zpravidla porovnává popisné statistiky souborů (průměry, směrodatné odchylky), míry reliability (např. Cronbachův koeficient alfa) obou verzí, korelační koeficienty a složení faktorů. V teorii odpovědi na položku se obecně porovnávají charakteristické (ICC) a informační (IIC) křivky položky v různých módech sběru dat. Dále se také zkoumá linearita vztahu mezi obtížností položky a její schopností diskriminovat, tedy mezi parametry *a* a *b* (Ferrando & Lorenza-Seva 2005: 10-11).

---

dotazování respondentů za použití počítače, CAPI (computer assisted personal interview) – osobní dotazování, při kterém se odpovědi respondentů zaznamenávají přímo do počítače.

[25] Výzkum CVVM SOÚ AV ČR, v.v.i., Naše společnost.

2. 1. 2. Variabilita dotazníků

Další výhodou teorie odpovědi na položku a jejího uplatnění v sociologii je možnost variability a flexibility dotazníku. V klasické teorii testů se nedají příliš dobře porovnávat dva výzkumy při změně jejich položek, např. při jeho zkrácení. Ale v IRT jsou individuální skóry nezávislé na setu otázek v dotazníku.

Výzkumy lze porovnávat pomocí velkého množství metod. Obecně jsou metody založeny na předpokladu, že (1) alespoň někteří respondenti zodpověděli všechny otázky u původního i nového testu nebo (2) některé položky jsou stejné v obou testech tzv. ukotvení položek

Možnost variability dotazníků je základní předpoklad pro počítačové adaptivní testování, které umožňuje efektivně měřit různé úrovně latentní proměnné a pomocí odlišného a personifikovaného testu dobře rozlišovat mezi respondenty (více viz kapitola 2. 4.

IRT se hodí pro vývoj standardizovaných testů, které se opakují v čase i v různém místě. Díky tomu, že odhadujeme položky a respondenty zvlášť, lze například určit velikosti latentní proměnné u respondentů okamžitě na základě určitého vzorce odpovědí.[26]

2. 2. Bias, odlišné fungování položek

Jedna z obtíží sociologických výzkumů, nejen mezinárodních, bývá tzv. bias. Bias neboli systematická chyba znamená, že naměřené hodnoty se odchylují od pravdivé hodnoty, a to nikoliv náhodně. Takovéto systematické chyby mohou vznikat v důsledku nevhodně zvoleného výběru souboru či neadekvátního zpracování dat

[26] Například pokud po řadě respondent odpoví na dotazník například ano, ne, ne, ano, ano, ne či velmi spokojen, spíše spokojen, spokojen, spíše nespokojen, velmi spokojen, můžeme, pokud už jsou položky předem kalibrovány, okamžitě určit měřenou latentní proměnnou i pro jednotlivce.

(Jandourek 2003: 206), či chyba způsobená návratností (Krejčí 2007). Další z mnoha důvodů mohou být právě kulturní rozdíly (Funkhouser 2001).

V řeči teorie odpovědi na položku můžeme systematickou chybu popsat jako situaci, kdy respondenti s obdobnou úrovní latentní proměnné odpovídají s navzájem odlišnou pravděpodobností na určitou položku. Například pro měření depresí v rámci MMPI – 2[27] položka „Snadno se rozbrečím.“, ukazuje, že ženy mají vyšší tendenci k depresím než muži (Reeve 2002: 38). Nicméně tato interpretace není dostatečná, neboť muži jsou jako skupina společensky nuceni neukazovat takovéto emoce. Položka „Snadno se rozbrečí“ měří tedy pro obě skupiny jinak. Muži a ženy se stejnou úrovní latentní proměnné, tedy se stejným sklonem k depresím, mají jinou pravděpodobnost, že na položku „Snadno se rozbrečím“ odpovědí kladně.

Problémy se systematickou chybou silně ohrožují validitu celého výzkumu, neboť v teorii odpovědi na položku se předpokládá, že se všechny položky „fungují“ stejně pro všechny respondenty (Johnson 2004: 19). Odhadujeme tedy univerzální parametry, které platí pro všechny skupiny, tedy v našem příkladu pro muže i pro ženy. Pokud by se parametry pro různé skupiny respondentů statisticky významně lišily, objevuje se ve výzkumu pravděpodobně nějaká (1) systematická chyba, bias anebo (2) skutečný rozdíl mezi skupinami.

V teorii odpovědi na položku se pro termín bias, systematická chyba, používá název odlišné fungování položky, *differential item functioning* (DIF). Pojem bias, systematická chyba, může být někdy zavádějící, neboť ne vždy je adekvátním označením situace. Někdy se nejedná o chybu v nástroji, ale právě o skutečný rozdíl mezi skupinami. Reeve (2002) argumentuje příkladem, kdy vědci studovali dva národy na stejné škále, a šlo jim o zjištění právě jazykových, nikoliv kulturních rozdílů. Stejně tak Ganglamair–Wooliscroft & Wooliscroft (2011) prováděli výzkum spokojenosti se

[27] MMPI – 2 je zkratka pro The Minnesota Multiphasic Personality Inventory (MMPI), což je jeden z nejpoužívanějších osobnostních testů pro měření duševního zdraví.

službami letecké společnost na různých národech. Cílem výzkumu bylo vyvinout škálu emocí, a to ve vícejazyčné verzi. Například slovo fantastický se na škále emocí v anglickém jazyce (fantastic) a v německém jazyce (fantastich) nacházelo jinde. Takže ačkoliv tato položka fungovala pro obě skupiny jinak, tak vzhledem k cíli výzkumu nelze mluvit o systematické chybě, ale správněji o odlišném fungování položky

Systematická zkreslení mohou být v kontextu teorie odpovědi na položku identifikována dvěma různými způsoby (Jelínek, Květoň & Vobořil 2011: 89). Jedním z nich je porovnání (1) parametrů pomoci klasického chí kvadrátu, případně porovnáním (2) plochy oblasti mezi charakteristickými křivkami položek[28]. V praxi je také důležitá vizuální analýza, porovnání tvarů funkce pro každou skupinu oddělené. Výsledky pak ukáží, zda jednotlivé položky měří pro každou ze skupin jiný konstrukt. Například DIF analýza odhalila v již zmíněném testu MMPI – 2, že položka „Během několika posledních let jsem se cítil(a) většinu času dobře“ byla hospitalizovanými respondenty zodpovídána ve smyslu fyzickém, zatímco ostatními respondenty ve smyslu psychickém, ve smyslu duševní pohody (Reeve 2002: 44). Stejná položka tedy u obou skupin měřila jiný konstrukt, jinou latentní proměnnou. DIF, odlišné fungování položek, se může zkoumat právě na úrovni jednotlivých položek, nebo i celého testu.

Teorie odpovědi na položku určitě zázračně nevyřeší veškeré problémy se systematickými chybami ve výzkumu, ani není jedinou možností, jak se s problematikou vypořádat, ale může významnou měrou přispět do aktuální sociologické praxe, zejména při mezinárodních výzkumech.

2. 3. Aplikace IRT v kognitivních přístupech

Kognitivní přístupy v sociologii se zabývají aspekty, které vstupují do procesu interakce tazatele a dotazované osoby s cílem odhalit zdroje zkreslení odpovědí (Vinopal

[28] Konkrétní matematické vyjádření lze běžně najít v knihách zabývající se IRT, například Jelínek, Květoň & Vobořil 2011: kapitola 7.

2008: 17). K problematice kvality otázek a kognitivním procesům lze přistupovat i moderními statistickými metodami jako je například teorie odpovědi na položku.

Kognitivní přístupy úzce souvisí s konstruktovou validitou (Embertson & Reise 2010), tedy jak kvalitně je daný teoretický konstrukt formulován do konkrétních položek v testu. Například závisí odpověď na položku více na porozumění otázce nebo na procesu rozhodování respondenta? Aplikace IRT v kognitivních přístupech je úzce spojena s multidimenzionální teorií odpovědi na položku, která je velmi zajímavá, ale také značně složitá a překračuje stanovený rámec diplomové práce.

Alespoň jako příklad z oblasti edukativních testů lze uvést logickou řadu obrázků, na jejíž konec se má doplnit správný obrázek. Než respondent dojde k úspěšnému výsledku, bylo pomocí kognitivní psychologie vysledováno, že musí projít třemi procesy. Jedním z procesů je využití (1) krátkodobé paměti, která může být kvantifikována počtem pravidel, kterými se řídí logická řada obrázků, dále závisí, (2) zda je vyžadována abstrakce a na (3) složitosti obrázku, zda je překlopen, překryt či nějak zkreslen. Na základě těchto třech procesů a jejich kvantifikace se pak vytvoří v rámci IRT patřičný model (Embertson & Reise 2010: 281).

Zda by takovéto modelování kognitivních procesů v rámci položky v sociologickém dotazníku našlo své uplatnění je vhodná výzkumná otázka pro další výzkumníky, kteří budou navazovat na tuto diplomovou práci.

2. 4. Adaptivní testování

Dotazník také může být zkrácen použitím tzv. adaptivního testování. Myšlenka adaptivního testování je poměrně prostá. Osoba zodpovídá jen takové otázky, které jsou pro měřenou úroveň latentního rysu nejrelevantnější. Položky s minimální informační hodnotou (např. příliš jednoduché otázky, kde by respondent udělal chybu pravděpodobně pouze z nepozornosti) nejsou tedy vůbec pokládány. Většina adaptivního testování je založena právě na teorii IRT (Rudner 1998).

Idea adaptivního testování se objevuje už na počátku 20. Století, kdy Alfred Binet sdružoval položky inteligenčních testů do skupin určených pro děti určitého věku (Jelínek, Květoň & Vobořil 2011: 104). S rozvojem informatiky a výpočetní techniky dochází k rozvoji i počítačového adaptivního testování, který celý proces zjednodušuje a automatizuje. Na základě zodpovězení každé předem zkalibrované položky v testu se znovu a znovu přepočítává úroveň měřeného latentního rysu jednotlivce, na jehož základě se pak vybírají další otázky tak, aby o každém respondentovi bylo získáno maximum možných informací.

Teorie odpovědi na položku slouží v rámci adaptivního testování právě ke kalibraci jednotlivých položek, k označení každé položky parametry. Na základě kalibrace se pak jednak vybírají nejrelevantnější položky dle aktuálního schématu odpovědí dotazovaného tak, aby byla změřena latentní proměnná co nejpřesněji. Ale také díky zkalibrovaným položkám je možné různé testy, které jsou obsahově odlišné, mezi sebou porovnávat. IRT tak slouží především v přípravné fázi výzkumu při vytváření tzv. „*item bank* / banky položek“ (Padaki & Natarajan 2009).

V komerční praxi se počítačové adaptivní testování používá například při skládání jazykových (TOEFL) či jiných (Microsoft softwarové certifikace) zkoušek (Květoň & Klimusová 2002: 252). Psychologický ústav AV ČR vyvinul vlastní program na počítačové adaptivní testování, který byl nazván CATo, Computerized Adaptive testing optimized. Byl úspěšně použit například při zkoumání neuroticismu (Jelínek, Květoň & Vobořil 2011: 117).

2. 5. Realita používání

I přes zjevné, v této druhé kapitole, zmiňované praktické aplikace a přínosy pro zkvalitňování metodologie výzkumů, se teorie odpovědi na položku na poli sociologie, alespoň v České republice, příliš nepoužívá (názor potvrzují i Urbánek & Šimeček 2001). Těžiště využití IRT stále zůstává v oblasti vzdělávání, kde se však používá i

u těch nejvýznamnějších mezinárodních šetření (například testy PISA[29] a jiné) a také v psychometrice.

Proč tomu tak je? Jeden z důvodů může být skutečnost, že výzkumníci byli učeni klasickou teorii testů, kterou znají a ve které umí interpretovat rozličné statistické proměnné (Reeve & Fayers 2005: 72). Stejně tak umí zacházet s příslušným programem, jako například SPSS. Naproti tomu IRT vyžaduje hlubší matematické schopnosti a příslušný software není tak rozšířený a „user friendly".

Dalším důvodem může být, že literatura a výzkumy využívající IRT se soustředí, jak již bylo zmíněno v úvodu této kapitoly, především na oblast vzdělávání. Výzkum, který sleduje například znalosti z dějepisu, tak může být sociologům vzdálený svým pojmoslovím, ale může být obtížné představit si použití takového nástroje na měření např. politické orientace.

Je také nutné připustit, že pravděpodobně ty nejužitečnější modely pro sociology, jako jsou polytomní modely pro data měřící postoje, jsou používány i ve světě stále zřídka (Drasgov 1990: 615). Příčinou této skutečnosti jsou stále ještě nedokonalé statistiky pro výzkum vhodnosti modelu a málo empirických šetření (opět Reeve 2002), natož nějaký standardizovaný postup.

Vždy také existuje časový rozdíl mezi novými metodami na úrovni teorie a jejich běžné aplikaci v praxi (Drasgow 1990). Často také výzkumníci nové metody neznají, nebo jsou ještě nekompletní či je hůře dostupný software. V devadesátých letech se dle Grasgowa (1990) začaly běžně používat jednodimenzionální dichotomní modely. V české sociologii ale ještě nedošlo ani k tomuto.

[29] PISA (Program for International Student Assessment) je uznávaný test, který slouží pro mezinárodní srovnávání žáků středních škol.

Přes všechna úskalí se začíná používání IRT rozšiřovat i do dalších disciplín jako je veřejné zdraví či psychologie. Především česká psychometrika se začala teorií odpovědi na položku v posledních letech zabývat a používat ji. Jedná se především o skupinu psychologů angažující se na Masarykově Univerzitě v Brně, jmenovitě například PhDr. Martin Jelínek, Ph. D. pod jehož vedením se i další studenti na univerzitě věnují IRT ve svých diplomových pracích (např. Kujal 2008). Je také autorem první české publikace věnující se systematicky teorii odpovědi na položku (Jelínek, Květoň & Vobořil 2011).

Jako další v České republice intenzivně využívá teorii odpovědi na položku společnost Scio (2008). Jako příklad lze uvést národní srovnávací zkoušky, které slouží také jako přijímací řízení na některé české vysoké školy. Sociologii bližší je nový počin společnosti Scio, která využívá teorii odpovědi na položku také v nástroji Personline (Scio 2010), což je systém personální diagnostiky, který je využíván především při náboru zaměstnanců.

Autorka věří, že se zvyšujícím se povědomí o teorii odpovědi na položku v České republice, a snad i s přispěním této diplomové práce, se i přes zjevná úskalí podaří i v české sociologii využívat přínosy IRT pro zkvalitňování výzkumů. Teorie odpovědi na položku je také svým způsobem mezioborová záležitost, která vyžaduje syntézu znalosti a spolupráci nejen sociologů, ale i statistiků, matematiků a vědců, zabývajících se psychometrií, kteří jsou alespoň v České republice metodologicky dále, dokonce i s vlastním programem pro počítačové adaptivní testování[30].

[30] Software CATo – Computer Addaptive Testing Optimized byl vyvinut na půdě Psychologického ústavu AV ČR, v.v.i.

3. Ilustrativní příklad

Využití teorie odpovědi na položku bude ilustrováno na výzkumu, jehož cílem je vyvinout univerzální nástroj pro měření subjektivní kvality pracovního života (Vinopal 2011). Na tento empirický nástroj, který byl vyvinut v rámci a v možnostech klasické teorie testů, bude pohlíženo právě z pohledu IRT. Teorie odpovědi na položku může do již existujícího nástroje přinést odpovědi na otázky jako:

(1) Jakou roli hraje každá položka v baterii? Neměří některé položky stejně?

(2) Rozlišují respondenti dobře odpovědi v nabízené škále?

(3) Jaký je celkový informační přínos každé položky pro výsledek testu?

(4) Pokrývá baterie celé kontinuum? Nebo měří lépe v některých jeho částech?

3. 1. Metodologie

Předložený nástroj (Vinopal 2011) se na subjektivní kvalitu pracovního života dívá ze dvou dimenzí, a to (1) spokojenosti a (2) důležitosti. Nástroj tedy obsahuje dvě baterie. Respondenti zhodnotí důležitost položek a následně spokojenost s nimi. Položky byly vybírány nejprve na základě teoretického konceptu a následně otestovány v pilotním šetření a redukovány na současnou podobu (více viz Vinopal 2011).

3. 1. 1. Data

Datový soubor pro tento ilustrativní příklad, jak už bylo popsáno v úvodu práce v kapitole 1. 3. 1, pochází z výběrového šetření „Naše společnost“, který byl realizován v roce 2009 Centrem pro výzkum veřejného mínění, SOÚ AV ČR pod názvem Stres na pracovišti – možnosti prevence. Na základě kvótního výběru bylo vybráno 950

zaměstnanců v České republice z věkové kategorie 18 až 65 let. Celkově bylo dotázáno 836 respondentů metodou standardizovaného rozhovoru s tazatelem na základě dotazníku.

Použitý nástroj pro výzkum subjektivní kvality pracovního života se skládá z 36 položek, které jsou rovnoměrně rozděleny do dvou baterií otázek pro (1) důležitost indikátorů a (2) spokojenosti s nimi. Odpovědi lze zvolit ze šestibodové škály „velmi spokojen – spokojen - spíše spokojen - spíše nespokojen - nespokojen, velmi nespokojen“, dále bylo možné zaznamenat odpověď „netýká se“ a „neví“. Obdobně pro baterii důležitosti.

Pro logickou interpretaci byla data překódována tak, aby vyšší spokojenost měla i vyšší číslo a to na hodnoty 1 (velmi nespokojen) až 6 (velmi spokojen). Použitý software IRTPRO je automaticky ještě následně překódoval na hodnoty 0 až 5 (viz např. Graf 11 až Graf 23). Zároveň odpovědi „netýká se“ a „neví“ byly považovány za tzv. missing values, aby mohl být použitý ordinální, nikoliv značně složitější nominální model. Vypovídající hodnota analýzy nebude ohrožena, protože pokud jsou chybějící hodnoty náhodné, nikoliv systematické, můžeme je pro tvorbu modelu zanedbat (Johnson 2004: 20).

Pro tuto diplomovou práci byla k analýze vybrána pouze baterie subjektivní spokojenosti (viz Příloha 1), nikoliv celý nástroj. Jedním z důvodů bylo ohrožení předpokladu unidimenzionality, ale i plánovaný rozsah práce. Nicméně dle názoru autora nástroje (Vinopal 2011: 948) i analýza každé dimenze samostatně je smysluplná a přináší zajímavé výstupy.

3. 1. 2. Použité nástroje

Baterie otázek na měření subjektivní spokojenosti s pracovním životem, která vznikla a byla analyzována v rámci CTT, bude zpracována optikou teorie odpovědi na položku. Překódování a úprava vstupních dat proběhla ve statistickém softwaru SPSS. Veškeré výpočty v rámci teorie odpovědi na položku byly zpracovány v programu IRTPRO

(Item Response Theory for Patient-Reported Outcomes), jehož verzi IRTPRO Student 2. 1. mohou studenti získat pro vědecké účely zdarma[31].

Pro finální úpravu grafů, především pro jejich názorné popisy, byl použit program Adobe Illustrator, neboť IRTPRO sice podporuje úpravu samotných grafů, ale není možné do nich nic vkládat ani vpisovat.

3. 1. 3. Předpoklady

Pro použití IRT jsou formulovány tři základná předpoklady a minimální velikost vzorku (více v kapitole 1. 4. které musí být pro správné použití teorie odpovědi na položku splněny.

(1) Unidimenzionalita byla zhodnocena pomocí tzv. scree plotu, který jednoznačně ukázal jednu dominantní dimenzi. Mezi první a druhou Eigenvalue je velký rozdíl (po řadě 5,4; 1,3 a 1,1). Pokud je splněn předpoklad unidimenzionality, je prakticky splněn i (2) předpoklad nezávislosti položek (Orlando 2004). Toto tvrzení je podpořeno i teoretickým konceptem a precizním sestavením dotazníku (více Vinopal 2011).

Základní otázkou pro smysluplné a správné použití IRT je (3) výběr modelu. Model musí odpovídat empirickým datům. Položky měřící subjektivní spokojenost s pracovním životem byly zodpovídány na šestibodové ordinální škále[32], proto byl zvolen polytomní unidimenzionální model odstupňovaných odpovědí (GRM), jehož autorkou je F. Samejima (1996). Vzhledem ke skutečnosti, že baterie je složena ze šesti

[31] IRTPRO Student je k dispozici na http://www.ssicentral.com/irt/student.html. Oproti plné placené verzi má několik omezení, co se týče především velikosti souboru a propojení s dalšími programy. Například v této verzi není možný přímý import dat z SPSS (sav). Tato na první pohled překážka se dá obejít uložením souboru v SPSS jako comma-separated values (csv). Tento formát už lze do programu IRTPRO Student bez problémů importovat.

[32] Kromě šesti nabízených odpovědí byla možná i varianta neví a netýká se. Zapojit tyto kategorie do modelu však přesahuje rámec této diplomové práce i možnosti diplomantky. Byly proto označeny jako chybějící odpovědi tzv. missing values.

domén se snahou měřit nejen celkovou spokojenost s pracovním životem, ale poskytovat i možnosti analýzy na úrovni jednotlivých položek a domén (Vinopal 2011: 955), byl by pravděpodobně vhodnější vícedimenzionální IRT model, který je však značně komplexnější, složitější a mimo stanovený rozsah diplomové práce. V tomto případě bude tedy dostačující postihnout pouze celkové subjektivní spokojenosti s pracovním životem.

Velikost vzorku s 836 respondenty, i pokud vezmeme v potaz chybějící odpovědi, je pro požadovanou analýzu dostatečná. Dle Embertson & Reise (2000) se u modelu odstupňovaných odpovědí (GRM) doporučuje pro přesné odhady parametrů velikost okolo 500 respondentů.

Analyzovány byly všechny tři varianty baterie měřící subjektivní spokojenost s pracovním životem, plná verze s 18 položkami a zkrácená s 12 a 6 položkami (více viz Vinopal 2011). Přehledné statistky celkové vhodnosti modelu baterie jsou prezentovány v následující tabulce (viz Tab. 3).

Tab. 3: Přehled vhodnosti modelu pro baterie o 6, 12 a 18 položkách

Statistiky	**18 položek**	**12 položek**	**6 položek**
RMSEA „model fit“	0,14	0,07	0,04
„items fit“/ items (α=0,05)	9/18	11/12	6/6
AIC	37 296,48	25 632, 48	13780, 15
BIC	37 807,18	25 972,94	13 950, 38
-2loglikelihood	37 080,48	25 488,48	13 708, 15

Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Často používanou statistikou pro ověřování shody modelu s odhadnutými parametry je index RMSEA (*root mean square of aproximation*). Čím více se RMSEA blíží nule, tím více model odpovídá realitě. Za hraniční je v posledních letech považována hodnota 0,07 (Hooper, Coughlan y Mullen 2008). Model je tedy shodný s daty pro zkrácené baterie o dvanácti a šesti položkách (viz Tab. 3). Jednodušší model vykazuje lepší shodu i dle dalších statistik (AIC, BIC a -2loglikelihood).

Druhý řádek z přehledu vhodnosti modelu (Tab. 3) popisuje shodu jednotlivých položek s modelem na základě statistiky chí kvadrát. Konkrétně vyjadřuje, kolik položek z celku

je ve shodě s modelem. Baterie s osmnácti položkami se neshoduje v polovině případů, zatímco ve zkrácené verzi se s modelem shodují, až na jednu výjimku, všechny položky. Na tomto místě je vhodné položit otázku, proč baterie s osmnácti položkami vykazuje špatnou shodu s modelem, zatímco její kratší verze se s modelem shodují?

Na vysvětlení (ne)shody dat s modelem bohužel neexistuje jednoznačná odpověď. Problém leží pravděpodobně u jednotlivých položek, neboť při analyzování kompletní baterie nebyla s modelem odstupňovaných odpovědí shodná polovina položek. Důvodem neshody může být například nerovnoměrné rozdělení odpovědí na škále. Když jsou například odpovědi na šestibodové škále kumulovány do jedné, dvou kategorií (jako například u položky „Jak jste spokojen s mírou výskytu násilí a šikany na pracovišti“, více viz kapitola 3. 2. 2), tak poté jsou parametry *b* nedostatečně odhadnuty. U některých položek (například „Jak jste spokojen s tím, kolik Vám práce umožňuje mít času na Vaši rodinu, zájmy a relaxaci?“) zůstávají některé z kategorií odpovědí nevyužity, neboť pro žádnou úroveň latentní proměnné není daná kategorii tou nejpravděpodobnější volbou. Možná by pak bylo vhodnější zvolit například jiný, nominální model, či pomocí dalších technik odhadnout precizněji parametry.

Obecně tématika vhodnosti modelu je stále předmětem aktuálních výzkumů a pro její hlubší analýzu už není v této diplomové práci prostor. Představuje ale jeden ze směrů, kam by se mohl dále ubírat výzkum a rozšíření této práce. Nyní bude pro další analýzy použit zkrácený instrument s dvanácti položkami (více Vinopal 2011 a Příloha 1), který uspokojivě odpovídá modelu, splňuje předpoklady a zároveň umožnuje charakterizovat více položek.

3. 2. Analýza a výstupy

Cílem této kapitoly je zanalyzovat baterii měřící subjektivní spokojenosti s pracovním životem pomocí teorie odpovědi na položku a dát tyto výstupy do souvislosti s klasickou teorií testů, v jejímž rámci byla baterie původně zpracována (Vinopal 2011). Snahou je tak ukázat užitečnost a rozdílnost, ale v některém případě i podobnost obou teorií a unikátně tak použít IRT v české sociologické studii. Pro následující kapitoly

bude až na výjimky použita zkrácená verze baterie o dvanácti položkách (důvody na konci předchozí kapitoly 3. 1. 3).

3. 2. 1. Obtížnost a diskriminace položek

Nejprve se podíváme na jednotlivé položky baterie v optice CTT a IRT. Následující tabulka (Tab. 4) ukazuje běžně používané statistiky klasické teorie testů, mezi které kromě (1) průměru patří (2) korelace položky s ostatními položkami a (3) Cronbachovo alfa (Sharkness & DeAngelo 2011: 489). Zároveň jsou přidány i standardní statistiky IRT – (4) parametr *a,* schopnost diskriminace položky, a (5) parametry *bi,* obtížnost položky. V následujících odstavcích bude provedena jejich interpretace a porovnání.

Tab. 4: Vlastnosti jednotlivých položek baterie v CTT a IRT

		CTT				IRT					
		1.		2.	3.	4.	5.				
	Položka	**prům**	**směr. odch.**	**Item -Tot. Corr**	***α* if item deleted**	**a**	**b1**	**b2**	**b3**	**b4**	**b5**
1a	Plat	2,80	1,13	0,60	0,88	1,68	-2,49	-1,52	-0,43	0,88	2,77
2b	Spravedlnost	2,86	1,13	0,65	0,87	2,06	-2,41	-1,49	-0,40	0,70	2,25
3d	S kolegy	3,72	1,01	0,57	0,88	1,63	-3,38	-2,71	-1,89	-0,49	1,27
4e	S nadřízenými	3,35	1,21	0,70	0,87	2,20	-2,43	-1,75	-1,00	-0,02	1,36
5g	Čas. náročnost	3,04	1,13	0,51	0,88	1,24	-3,12	-2,24	-0,76	0,50	2,72
6h	Rozložení pr. doby	3,18	1,14	0,52	0,88	1,33	-3,10	-2,19	-0,98	0,30	2,16
7j	Zajímavost	3,43	1,15	0,64	0,87	1,78	-3,15	-2,10	-1,19	0,01	1,04
8k	Vzdělání	3,03	1,21	0,65	0,87	1,85	-2,63	-1,60	-0,67	0,49	1,85
9m	Typ prac. poměru	4,13	1,02	0,52	0,88	1,29	-4,12	-3,38	-2,51	-1,35	0,38
10n	Jistota místa	3,23	1,29	0,55	0,88	1,36	-2,95	-2,01	-1,02	0,05	1,75
11p	Bezpečnost	3,74	1,00	0,60	0,88	1,73	-3,75	-2,74	-1,74	-0,42	1,13
12r	Čistota	3,73	1,05	0,54	0,88	1,43	-3,81	-2,88	-1,90	-0,50	1,14

Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Cronbachova alfa je v CTT často používaná statistika měřící míru interní konzistence, která je žádoucí pro vytvoření reliabilního nástroje. Tato baterie vykazuje velmi dobrou interní konzistenci. Statistika (3) Cronbachova alfa při vynechání položky (třetí sloupec „*alfa if item deleted*" v Tab. 4) ukazuje, jak se změní interní konzistence, pokud položku do baterie nezařadíme. (2) Korelace položky s ostatními položkami (*item total correlation*) ukazuje vztah mezi jednotlivou položkou a celkovou baterií. V baterii subjektivní spokojenosti s prací se korelace pohybuje od hodnoty 0, 51 až 0, 70

a ukazuje se, že každá položka z baterie je s ní koherentní a přispívá k měření celkového konstruktu. Na základě těchto statistik a znalosti pozadí výzkumu (Vinopal 2011) můžeme shrnout, že v optice CTT se jedná o reliabilní a kvalitní baterii.

Ve škálách obecně nejsou všechny položky asociovány s měřenou latentní proměnnou stejně. Relevanci každé položky v klasické teorii testů zkoumáme, jak již bylo zmíněno, například její korelací s celkovým skórem testu (*item total correlation*). V IRT je tato charakteristika zastoupena (4) parametrem *a*, který tak vlastně ukazuje váhu každé položky. Hodnoty parametru *a*, schopnost diskriminace položky, jsou nad hodnotu 1, 70[33] považovány za velmi vysoké, což je žádoucí. Hodnoty mezi 1, 35 a 1, 70 jsou považovány za vysoké a ty mezi 0, 65 a 1, 34 jako průměrně diskriminující (Baker 2001).

Parametr *a* (viz Tab. 4) dosahuje v našem příkladu od velmi vysokých hodnot 2, 20 až po relativně nízké ale stále dostatečně silné 1, 24. Diskriminace většiny položek patří mezi vysoké či velmi vysoké, což znamená, že většina položek přispívá relativně velkým dílem do měření subjektivní spokojenosti s pracovním životem. Pokud sestavíme pořadí, dle korelace *r* či parametru *a*, nepozorujeme příliš zásadní rozdíly mezi oběma teoriemi. Otázka na spokojenost s chováním nadřízených (položka 4, r = 0, 70 resp. a = 2,20) patří mezi nejvíce užitečné otázky pro měření konstruktu. Naopak nejméně podle obou teorií přispívá spokojenost s časovou náročností práce (položka 5, r = 0,51, resp. a = 1, 24).

Při analýze položek v modelu není důležitá pouze jejich relevance, respektive diskriminace, ale také jejich „obtížnost“ či jejich umístění na kontinuu. Zjišťujeme tak například, zda není test z matematiky příliš lehký, či zda neměříme naší baterií pouze ty ne tolik spokojené respondenty. V IRT obtížnost položky představují (5) parametry b. Položky v baterii mají šest kategorií, tedy u každé se odhaduje 5 parametrů obtížnosti *b1* až *b5*. Interpretace parametru *b* například pro položku plat (*b1*= - 2, 49*,* viz Tab. 4)

[33] Všechny hodnoty diskriminačního parametru *a* musí být interpretovány v logistické metrice.

vypadá následovně: Model předpovídá, že respondent se subjektivní spokojeností s prací 2, 49 směrodatné odchylky pod průměrem (θ = - 2, 49) má 50% šanci, že na otázku „Jak jste spokojen s výší platu nebo mzdy“ odpoví „velmi nespokojen“.

Jak je vidět z hodnot parametrů *b* z Tab. 4 je jich většina negativních (41 paramtrů *bi* ze 60). Hraniční hodnoty b1 pro nejvíce negativní odpověď (velmi nespokojen) se pohybují od - 4,12 do - 2, 41. Naopak pro *b5* od 0, 38 do 2, 77. Tyto údaje ukazují, že baterie pravděpodobně lépe pokrývá negativní stranu kontinua, tedy tu podprůměrnou spokojenost s pracovním životem. Z této skutečnosti plyne, že ti, kteří jsou velmi spokojeni, mají k dispozici jen málo možností, jak se vyjádřit, a škála rozlišuje lépe mezi méně spokojenými respondenty. Toto tvrzení bude dále analyzováno graficky v kapitole 3. 2. 2).

Alternativou k parametru *b* je v klasické teorii testů obtížnost položky měřená například (1) průměrem, či jinou statistikou polohy. Z Tab. 4 je zřejmé, že nejvyšší průměr byl zaznamenán pro spokojenost s charakterem pracovního poměru (položka 9, průměr = 4, 13), což odpovídá kategorii spíše spokojen. Všechny parametry *b* u položky 9 jsou záporné nebo velmi nízké, což indikuje, že položka patří mezi ty „lehké“.

Lze tedy shrnout, že IRT i CTT poskytují do jisté míry obdobné informace o jednotlivých položkách a nejsou v rozporu. Nicméně IRT nám poskytuje ještě další možnosti, jak interpretovat či vytvářet měřící nástroje.

3. 2. 2. Kategorie odpovědí na položky

V následujících kapitolách přejdeme od numerického popisu modelů ke grafickému. Ačkoliv běžně dostupný software pro provádění IRT analýz není příliš uživatelsky přívětivý a nedosahuje kvalit a přehlednosti programů, které slouží pro klasickou teorii testů (SPSS, SAS), tak grafické výstupy jsou velmi názorné a intuitivní pro širokou škálu uživatelů.

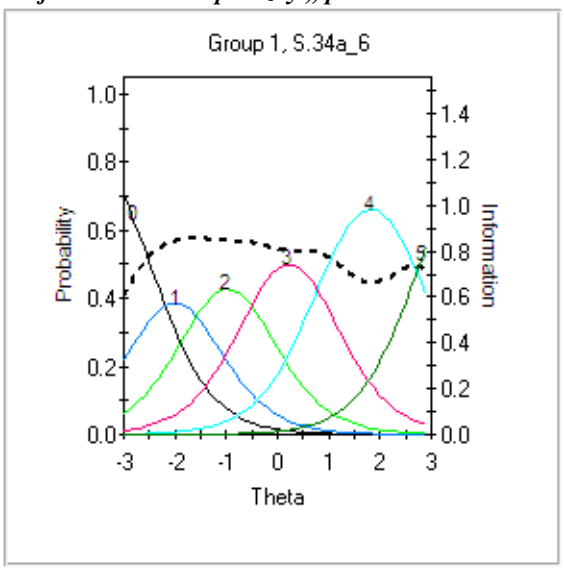
Následující grafické výstupy lze analyzovat ze dvou základních pohledů, a to (1) rozložení pravděpodobnosti zvolení určité kategorie na kontinuu, kterou představují charakteristické křivky položky (ICC). Druhý pohled představuje (2) informační funkce položky (IIC), kterému se budeme věnovat v následující kapitole (kap. 3. 2. 3).

Při vývoji dotazníku je jednou ze základních otázek, jaké zvolit kategorie odpovědí, zda jednoduché dichotomní ano/ne, pětibodovou, sedmibodovou či jinou škálu. Pokud jsou k dispozici data z pilotního projektu, je možné určit jejich vhodný počet. V klasické teorii testů se lze podívat na frekvenci odpovědí. To však může být matoucí, neb i ty nejméně obsazené kategorie mohou přinášet velmi důležité informace právě tím, že pokrývají okrajovou část kontinua.

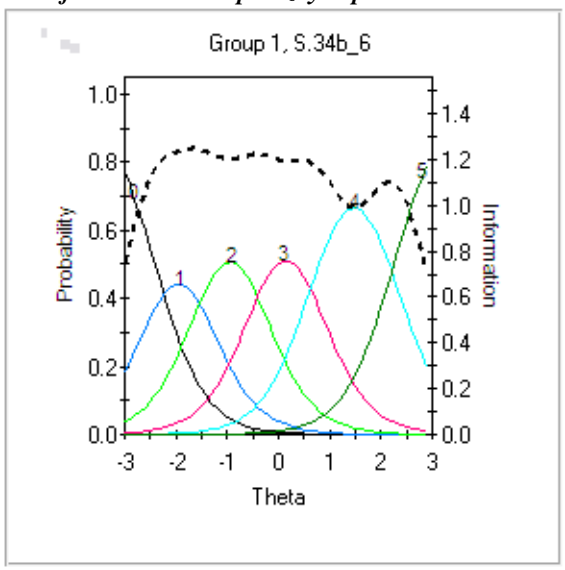
Na následujících grafech (Graf 11 až Graf 22), které zobrazují vztah subjektivní spokojenosti s prací (theta) a pravděpodobnosti, s jakou respondent odpoví určitou kategorií, lze velmi dobře pozorovat právě rozložení jednotlivých kategorií[34]. Tato vizualizace tak efektivně doplní informace již známé z odhadnutých parametrů, popsaných numericky v předchozí kapitole (Tab. 4). Například u položky „plat“ (viz Graf 11) jsou kategorie poměrně pravidelně rozložené. Zatímco u položky spokojenosti se vztahy s kolegy (Graf 13), je zjevná kumulace v levé části kontinua. Kategorie nespokojen (1) je překryta ostatními kategoriemi. Odpověď „nespokojen“ nemá při žádné úrovni subjektivní spokojenosti s prací nejvyšší pravděpodobnost, že jí někdo použije, proto by mohla být teoreticky vyloučena.

[34] V použitém software ještě není implementována úplně snadná úprava grafů. Odpovědní kategorie na grafech jsou tedy: (0) velmi nespokojen, (1) nespokojen, (2) spíše nespokojen, (3) spíše spokojen, (4) spokojen a (5) velmi spokojen.

Graf 11: ICC a IIC položky „ plat“

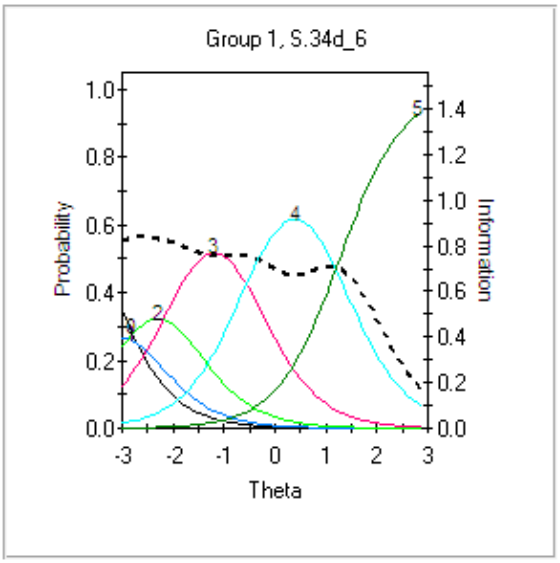


Graf 12: ICC a IIC položky "spravedlnost"

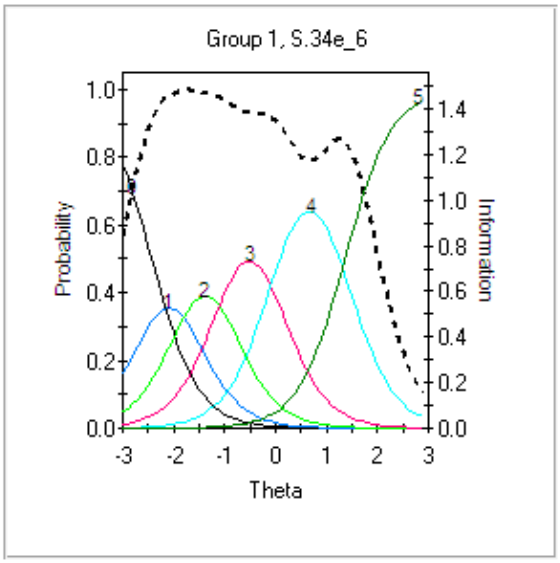


Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Graf 13: ICC a IIC položky „s kolegy“



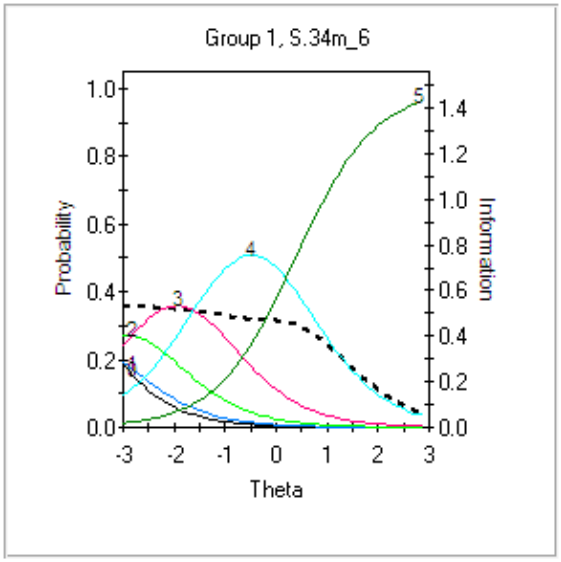
Graf 14: ICC a IIC položky „s nadřízenými“



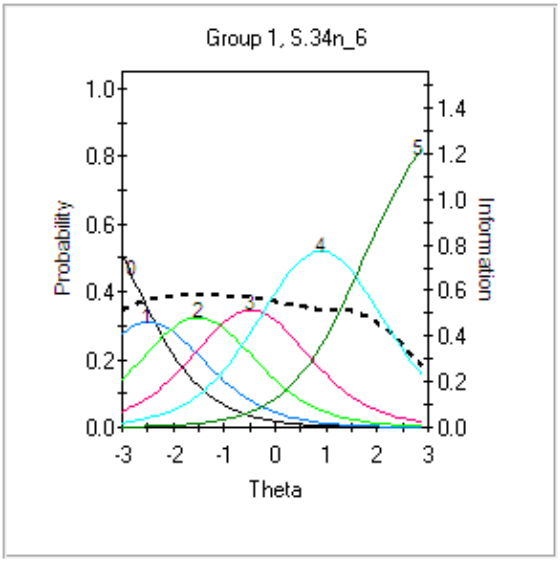
Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

U spokojenosti s typem pracovního poměru (Graf 15) lze zase pozorovat, že respondenti s už jen trochu nadprůměrnou spokojeností odpovídají s velmi vysokou pravděpodobností extrémní kategorií velmi spokojen. Proto tato položka dobře nerozlišuje mezi respondenty, kteří jsou spokojeni se svým pracovním životem, neboť všichni spadají do jedné kategorie odpovědi. U takovýchto položek je naopak teoreticky vhodné zvážit změnu počtu kategorií (Reeve & Fayers 2005) či upravit formulaci otázky, protože pro tuto položku není vhodných šest kategorií.

Graf 15: ICC a IIC položky „typ prac. poměru“



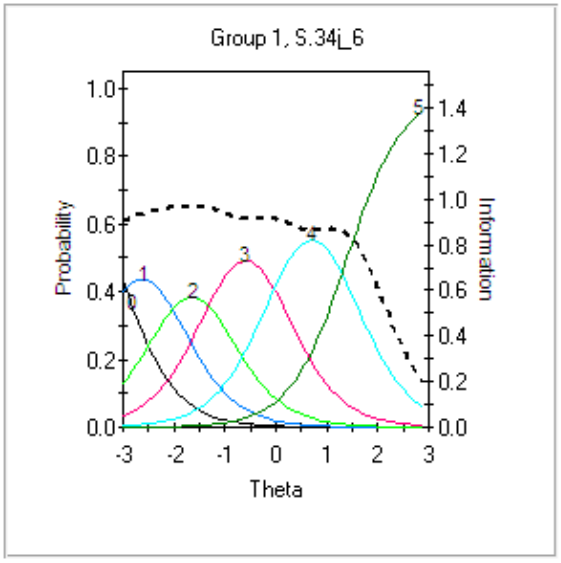
Graf 16: ICC a IIC položky „jistota místa“



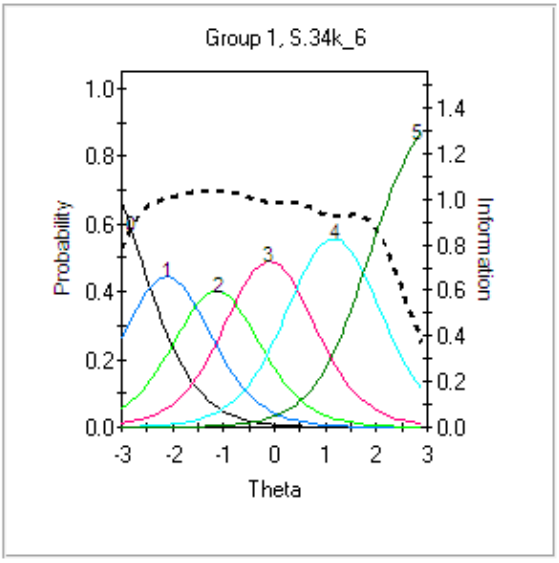
Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Některé z křivek odpovědí vypadají zase velmi podobně, až změnitelně, jako například položky z domény seberealizace, konkrétně „zajímavost práce“ (viz Graf 17) a „vzdělávání a osobní rozvoj“ (viz Graf 18). Vzorec odpovědí na takové položky je velmi podobný. Respondenti zvolí s vysokou pravděpodobnosti stejnou odpověď, ať už se jich zeptáme na spokojenost se zajímavostí práce či s možnostmi dalšího vzdělávání a osobního rozvoje.

Graf 17: ICC a IIC položky „zajímavost“



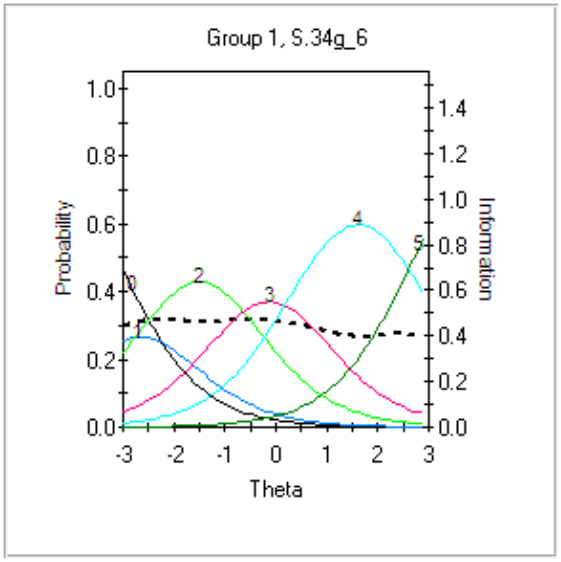
Graf 18: ICC a IIC položky „vzdělávání“



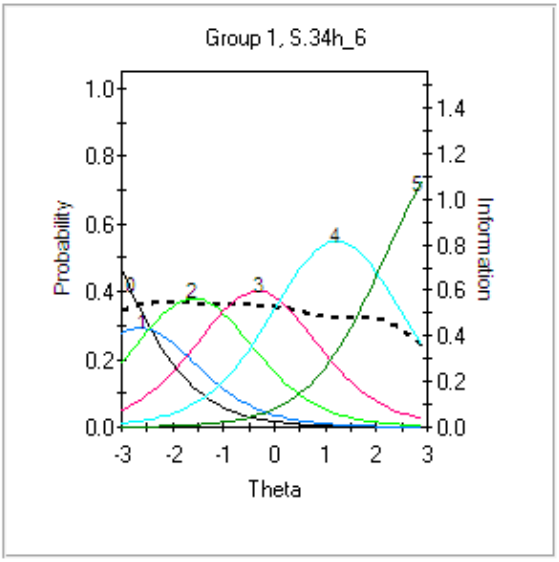
Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Obdobně můžeme analyzovat i další položky z baterie subjektivní spokojenosti s pracovním životem, položky z domény čas (viz Graf 19 a 20) a z domény podmínky (Graf 21a 22).

Graf 19: ICC a IIC položky „časová náročnost“

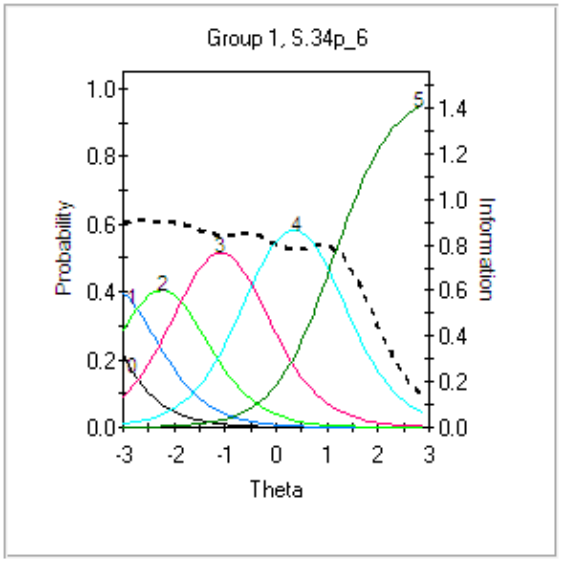


Graf 20: ICC a IIC položky „rozložení pr. doby“

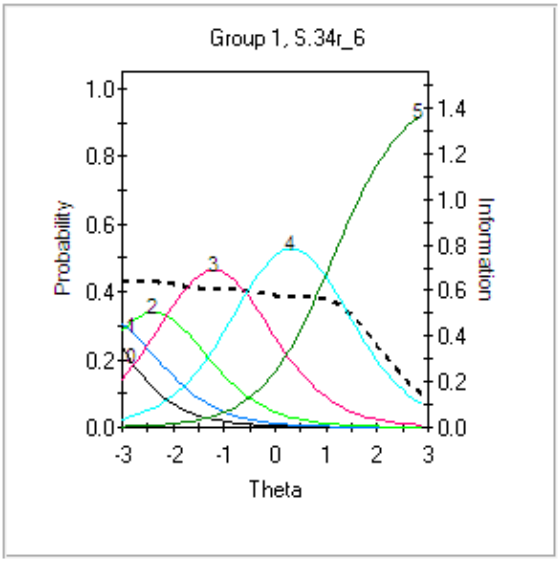


Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Graf 21: ICC a IIC položky „bezpečnost“



Graf 22: ICC a IIC položky „čistota“



Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Na základě vizuální analýzy jsme diskutovali několik možných výstupů, ať už konstatování o pravidelném rozložení kategorií (doména plat), doporučení o ubrání kategorie odpovědí (doména vztahy), či vůbec o revizi otázky a jejího počtu kategorií

(položka spokojenost s charakterem pracovního poměru) a dále možnost vynechání jedné z otázek pro velmi podobné, až zaměnitelné vzorce odpovědí (doména seberealizace).

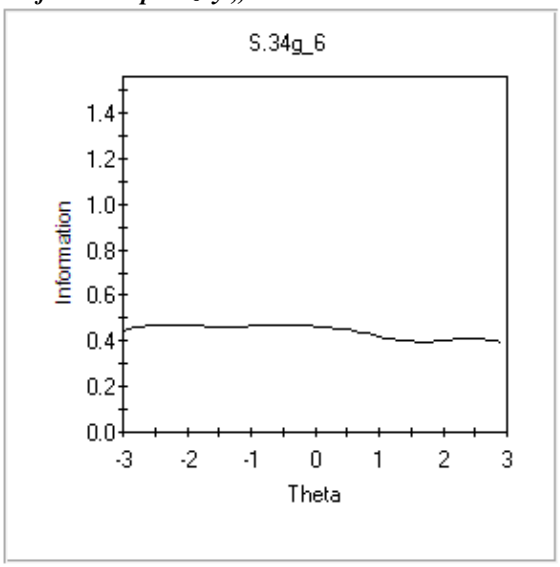
Nicméně takovéto změny v baterii nelze aplikovat jednoduše, neboť se jedná o ucelený nástroj. Pravděpodobně vylepšení každé jednotlivé položky a optimalizace jejího počtu kategorií by vedla ke zmatení. Respondenti by v rámci jedné baterie odpovídali pokaždé na jiné škále, což by mohlo ovlivnit jejich odpovědi a průběh dotazování, a negativně tak i kvalitu celého nástroje. Konečná úprava vždy záleží na konkrétní situaci a účelu výzkumu a úsudku výzkumníka uspokojivě vybalancovat požadavky statistické a praktické.

3. 2. 3. Informační přínos položek

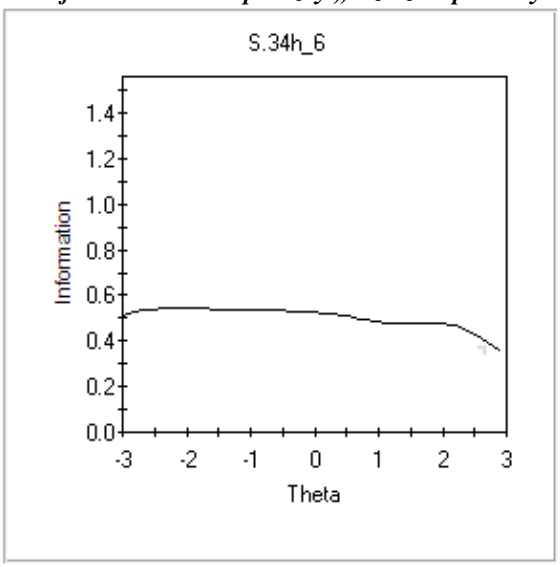
Informační funkce položky (IIC) ukazuje, jak velkou informaci každá jednotlivá položka přináší celému měřícímu nástroji a zároveň na jaké části kontinua se nejvíce realizuje. Informační funkce všech položek jsou vykresleny na předchozích grafech přerušovanou čárou (Graf 11 až Graf 22). Jejich analýza přispívá k optimalizaci a zvyšování reliability výzkumných nástrojů.

Díky IRT lze například zkracovat dotazníky při zachování reliability tak, že se vyloučí položky, které jsou sice odlišně formulované, ale v podstatě měří to samé. Položky, které v našem případě patří do jedné subdimenze, například do domény čas (viz přehlednější Graf 23 a 24), mají vysokou interní konzistenci a zároveň mají podobný průběh jak charakteristické (ICC), tak informační křivky (IIC) položky (viz Graf 19 a 20). Proto bychom ztratili jen málo informací, pokud by se jedna z položek vynechala a dotazník se zkrátil.

Graf 23: IIC položky „časová náročnost“



Graf 24: ICC a IIC položky „rozložení pr. doby“



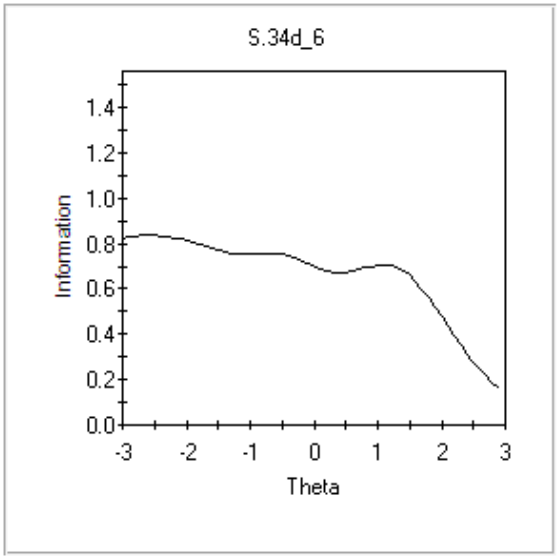
Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

Můžeme tedy systematizovat, že z informační křivky položky získáme informaci (1) jakou měrou přispívá k celkové informační hodnotě baterie. Pokud bychom chtěli nástroj zkracovat a do dotazníku zařadit jen jednu položku z domény například vztahy, byla by to rozhodně položka spokojenost s nadřízenými (viz Graf 26), která má největší informační přínos ze všech položek.

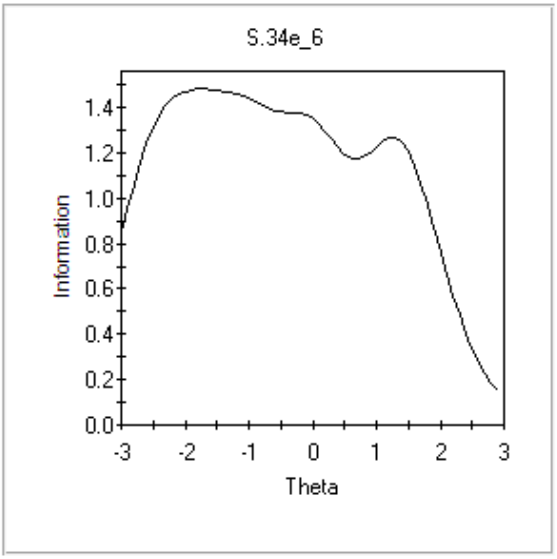
Dalším pohledem na informační funkci položky je (2) její průběh, například položky z domény čas (Graf 23 a 24) měří ve srovnání s položkami v doméně vztahy (Graf 25 a 26) poměrně konstantně v celé délce kontinua. Důležité je také (3) srovnání s ostatními položkami, aby výsledný nástroj, co nejlépe[35] pokrýval celé kontinuum a přinášel maximum informací.

[35] Nejlépe pokrývat kontinuum neznamená jen plochá informační křivka, ale taková, která nejlépe odpovídá našemu výzkumnému cíli. Pokud chceme vyvinout univerzální nástroj, je vhodná plochá křivka, pokud se však chceme zaměřit například na osoby nespokojené s prací, je žádoucí, aby právě v té části kontinua měřila položka nejlépe.

Graf 25: IIC položky „s kolegy“



Graf 26: IIC položky „s nadřízenými"



Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

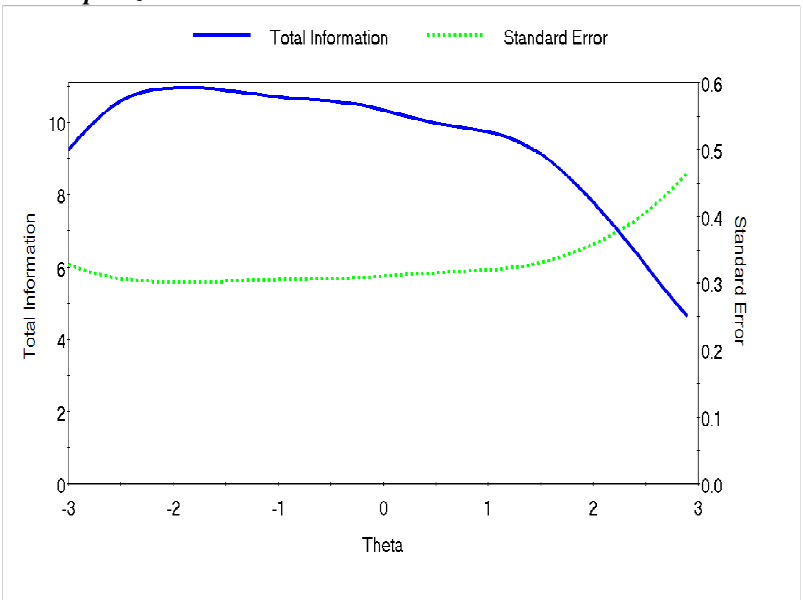
Vzhledem ke skutečnosti, že používaný nástroj na měření subjektivní spokojenosti s pracovním životem je určený i pro hodnocení jednotlivých aspektů, resp. domén, je na místě otázka, jak moc je zkracování baterie žádoucí. Zda bychom pak mohli oprávněně mluvit o spokojenosti v jednotlivých dílčích doménách, pokud by byly měřeny jen jednou či dvěma položkami. Na tomto místě je třeba upozornit, že je v rámci této diplomové práce zpracovávána pouze unidimenzionální IRT. Existují i vícedimenzionální modely, které by byly pro takovouto komplexní baterii, která by dobře fungovala i v jednotlivých aspektech spokojenosti jistě vhodnější, ale také výpočtově složitější.

3. 2. 4. Informační funkce baterie jako celku

V klasické teorii testů se pro hodnocení kvality baterie používají statistiky pro měření reliability. Míry pro vnitřní konzistenci jsou často zastoupeny statistikou Cronbachovo alfa. To, že má nástroj vysokou reliabilitu (viz kapitola 3. 2. 1), implikuje nízkou směrodatnou odchylku. A právě zde se projevuje jedna z výhod IRT. Na grafu celkové informační funkce baterie (viz Graf 27) lze sledovat měnící se hodnotu směrodatné odchylky, která v teorii odpovědi na položku není konstantní po celém měřeném kontinuu.

Nejvyšší reliabilitu vykazuje měřený instrument pro nespokojené osoby, tedy ty, které jsou na levé straně kontinua. Pro ty extrémně spokojené je celková vhodnost a preciznost testu až o polovinu menší (viz Graf 27). Logicky zrcadlově se pak chová směrodatná odchylka, která je tak nejvyšší pro osoby, které mají nejvyšší subjektivní spokojenosti s prací. Takováto užitečná informace jde za rámec klasické teorie testů a dobře ukazuje, že se obě teorie mohou výborně doplňovat a obohacovat. Nejvíce informace dostáváme pro respondenty s úrovní spokojenosti dvě směrodatné odchylky pod střední hodnotou. Při hodnotě latentní proměnné 1, 5 směrodatné odchylky nad průměrem začíná informační funkce poměrně ostře klesat a měřit precizně.

Graf 27: Celková informační funkce a směrodatná odchylka baterie dvanácti položek



Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

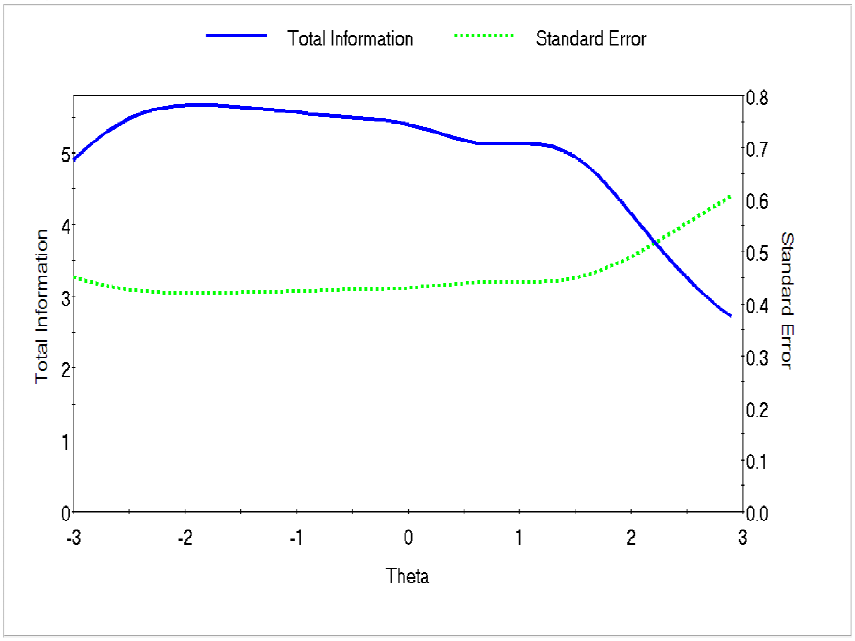
Například pro baterii, kterou bychom v rámci CTT a jejích statistik zhodnotili jako velmi kvalitní, lze v rámci IRT najít prostor pro vylepšení (více v kapitole 3. 2. 4). V CTT vnitřní konzistence, Cronbachovo alfa, roste s přidáním další položky do dotazníku. A to i tehdy, když je velmi podobná ostatním položkám, které už jsou v dotazníku obsaženy. Přidání takovéto položky však neposkytne příliš mnoho informací navíc a položka je v podstatě nadbytečná. Oproti tomu výstupy z IRT ukazují, jaká by položka měla být, zda těžší, či lehčí, na jaké části kontinua. Přidání takové položky pak zvyšuje nejen reliabilitu, ale i validitu celého nástroje (Sharkness & DeAngelo 2011). Pro univerzální měření spokojenosti s pracovním životem by bylo

vhodné zařazení takové položky / takových položek, které budou dobře diskriminovat, právě mezi těmi respondenty s vyšší spokojeností.

Nalezení vhodné doplňující položky pro větší preciznost měřícího nástroje vyžaduje důkladné zvážení toho, co znamená být velmi spokojen s pracovním životem a zformulovat otázky vhodné právě pro takovou cílovou skupinu. Objevením nové položky se nezvýší jen kvalita nástroje, ale také může přispět k vytváření teorie zabývající se oblastí spokojenosti s pracovním životem (Sharkness a DeAngelo 2011).

V použitém software IRTPRO lze interaktivně přidávat a odebírat položky a okamžitě se vizualizuje jejich přínos a jejich vliv na celkovou informační funkci. Vidíme, že šestibodová baterie (viz Graf 28) má logicky menší informační přínos a větší směrodatnou odchylku, ale velmi podobný průběh jako s dvanácti položkami.

Graf 28: Celková informační funkce a směrodatná odchylka baterie šesti položek



Zdroj: Vlastní zpracování na základě výpočtů v IRTPRO

3. 3. Diskuze

Pokud shrneme předchozí analýzy, můžeme říci, že v ohledu hodnocení jednotlivých položek baterie přináší IRT i CTT obdobné výsledky. Například otázka na spokojenost

s chováním nadřízených patří mezi nejvíce užitečné otázky pro měření konstruktu, naopak nejméně podle obou teorií přispívá spokojenost s časovou náročností

Co ale nemůžeme získat v klasické teorii testů, je zhodnocení, jak dobře měří škála podél kontinua. V našem příkladu měří instrument dobře spíše respondenty méně a středně spokojené. Ale není příliš precizní pro ty opravdu spokojené, není schopen mezi nimi uspokojivě rozlišovat. Tento pohled v CTT chybí, neboť kvalita instrumentu se hodnotila pomocí statistiky Cronbachovo alfa, které bylo velmi vysoké, a implikovalo nízkou, podél kontinua konstantní směrodatnou odchylku, což je znak kvalitního a konzistentního a reliabilního měřícího nástroje. V rámci IRT, kde není směrodatná odchylka konstantní, byl objeven prostor pro optimalizaci nástroje. Kromě možné diskuze o vynechání některých položek, či změnách kategorií odpovědí, je zde prostor pro přidání další položky, dimenze, domény, která bude dobře měřit respondenty velmi spokojené s prací. Autorka předpokládá, že by se další výzkum mohl ubírat cestou „svobody v práci“, což je přístup založený na tom, aby lidé v práci dělali to, co je baví a v čem jsou nejlepší (Carney & Getz 2011).

Myšlenku o doplňující doméně svobody, štěstí v práci nebudeme dále rozvíjet. Tato práce si neklade za cíl hodnotit baterii, měnit jí a zabývat se tématem subjektivní kvality pracovního života. Nesnaží se o interpretace a hledání odpovědí na nové otázky, které vyvstaly po provedení IRT analýzy, jako proč jsou odpovědi kumulovány na jedné straně kontinua, proč ta která položka přináší více informace, jaká dimenze by dále doplnila a zvýšila reliabilitu a validitu baterie. Nesnaží se ani formulovat doporučení a závěry. Snahou bylo spíše na použitém nástroji ukázat možnosti a aplikace IRT na sociologicky blízkém tématu.

Práce také ukázala velký prostor pro další výzkum a rozvoj, ať už na úrovni konkrétního nástroje, tak na úrovni aplikace IRT v sociologii. Velmi přínosné pro vytvoření univerzálního nástroje měřící subjektivní spokojenost s prací by také byla analýza odlišného fungování položek (DIF), která by například mohla odhalit, zda jednotlivé položky měří stejně pro manuální a nemanuální zaměstnance a osoby samostatně

výdělečně činné či pro jiné skupiny. Potřebným rozšířením by také byla aplikace vícedimenzionálního modelu, který by umožnil optikou IRT postihnout i jednotlivé domény baterie a umožnil tak jejich podrobnější analýzu a kvalitní sestavení konstruktu. Vícedimenzionální modely pro polytomní položek jsou však značně složitější a komplexnější a přesahují stanovený rámec diplomové práce.

Dále tato práce poskytuje jen pravděpodobný návrh, proč původní baterie o osmnácti položkách neodpovídá modelu, zatímco obě její kratší verze ano. Zajímavá by také v tomto kontextu byla analýza prvotní verze dotazníku, která obsahovala 66 položek pro spokojenost, a v optice IRT ji zhodnotit a optimalizovat a následně porovnat s nástrojem vytvořeným podle klasické teorie testů. To byl jen namátkový výčet témat, které nejsou v této diplomové práci obsažena, a která stojí za úvahu. Jistě by se našlo mnoho dalších možností a způsobů, kterými by se mohl ubírat další výzkum, nicméně tato diplomová práce má jen omezený rozsah.

Závěr

Ačkoliv je teorie odpovědi na položku stále relativně novou a vyvíjející se oblastí a ačkoliv byla doposud v České republice využívána především ve vzdělávacích a osobnostních testech, tak její zapojení do sociologické metodologie vneslo do kvantitativního výzkumu nový a velmi užitečný pohled, který pomáhá při sestavování krátkých, cílených a reliabilních instrumentů.

Možnosti teorie odpovědi na položku se postupně rozšiřují, nejen co se týče rozvoje teorie, ale také aplikací. Objevuje se více programů na zpracování dat v rámci IRT, více případových studiích s různorodou tématikou a také se kromě vyvíjení stále nových modelů a statistik začínají některé postupy už usazovat. Vznikající konsenzus pro používání základních statistik a modelů usnadňuje a vůbec i umožňuje práci širšímu spektru odborníků, kteří nejsou primárně zaměření na oblast teorie odpovědi na položku, ale spíše na její využití.

V této diplomové práci byly kromě zpracování teoretického základu IRT v jazyku sociologie formulovány možnosti uplatnění v sociologii. Mezi základní aplikace patří využití IRT pro vytváření krátkých a reliabilních výzkumných nástrojů díky možnostem modelovat charakteristickou a informační funkci položky i celé baterie. IRT umožňuje zjistit, kde je třeba precizněji zpracovat měřící nástroj, poskytuje rozložení položek dotazníku a respondentů na jedné škále a na základě modelu umožnuje zvolit vhodný počet kategorií odpovědí.

Vlastnosti položek IRT mají i další využití, například při porovnání reliability při různých módech sběru dat či při srovnatelnosti výzkumů při změně některých položek,

což je obzvláště užitečné pro opakující se výzkumy. Pomáhá také odhalovat systematickou chybu ve výzkumu a odlišné fungování položky pro různé skupiny. Dalšími možnostmi, kde se se IRT používá, je adaptivní testování, které připraví dotazník respondentovi přímo na míru, či kognitivní přístupy, které se snaží rozklíčovat vlivy vstupující do procesu interakce mezi respondentem a tazatelem.

I přes zjevné výhody a možnosti využití je IRT na poli české sociologie stále nevyužívaným a opomíjeným konceptem. Ve srovnání s CTT přispívá IRT významnou měrou do oblasti (1) reliability, kde se směrodatná odchylka na rozdíl od CTT liší pro různou úroveň latentní proměnné. Klíčovou výhodou IRT na úrovni položek je (2) jejich nezávislost na kontextu testu i výběrového vzorku. IRT také (3) nepožaduje normalitu dat a reprezentativnost výběru.

Teorie odpovědi na položku nemá ambice nahrazovat klasickou teorii testů v sociologii, či jí konkurovat. Jde spíše o potřebný a přínosný doplněk k sociologické metodologii, jak bylo ukázáno i na ilustrativním příkladu, který byl zpracován v rámci klasické teorie testů (Vinopal 2011). Optika IRT může nástroj pomoci dále vylepšit především zjištěním, že pomocí baterie měřící subjektivní spokojenost s pracovním životem není možné precizně změřit respondenty, kteří jsou nadprůměrně spokojení.

Diplomová práce přináší v českém jazyce pravděpodobně ojedinělé stručné shrnutí problematiky teorie a aplikace teorie odpovědi na položku v sociologii, a její následnou demonstraci na výzkumném nástroji měřící subjektivní spokojenost s pracovním životem. Celá práce je doplněna i o srovnání s klasickou teorií testů, úvahy o (ne) používání IRT v české sociologii a o inspiraci pro další výzkum, ke kterému doufáme, přispěje právě i tento text. Na závěr ještě považujeme za nezbytné zmínit, že byla zpracována pouze unidimenzionální IRT a zajisté nebyly využity všechny její možnosti, výhody a aplikace v celé šíři. Avšak i v této fázi práce představuje komplexní a do jisté míry originální studii, která kromě vlastních formulovaných závěrů přináší i návod a širokou inspiraci pro další výzkumníky, kteří mohou na práci „Teorie odpovědi na položku a její aplikace v sociologii“ navazovat.

SEZNAM POUŽITÉ LITERATURY

BAKER, F. 2001. *The Basics of Item Response Theory*. University of Maryland, MD: ERIC Clearinghouse on Assessment and Evaluation, 2001. [cit. 2012-08-07]. Dostupný z WWW: <http://echo.edres.org:8080/irt/baker>.

BOCK, R. D. 1997. A Brief History if Item Theory Response. *Educational Measurement: Issues and Practice.* 1997, vol. 16., no. 4, s 21 – 33. Dostupný také z WWW: < http://brainimaging.waisman.wisc.edu/~perlman/Kristin/bock-educ-meas-1997.pdf>.

BOGARDUS, E. S. 1933. A Social Distance Scale. *Sociology and Social Research*, 1933. vol 17., s. 265-271.

CARNEY, B. M. – GETZ, I. 2011. *Svoboda v práci*. Praha: Peoplecomm, 2011. 344 s.

DEMARS, CH. 2010. *Item Response Theory : Understanding Statistics Measurement*. USA: Oxford University Press, 2010. 131 s. ISBN 9780195377033.

DRASGOW, F. – HULIN C. L. 1990. Item Response Theory. In M. D. DUNNETTE – L. M. HOUGH (Eds*.). Handbook of Industrial and Organizational Psychology*. Palo Alto: Consulting Psychologists Press, 1990. s 577- 636. Dostupný také z WWW: <http://mres.gmu.edu/readings/PSYC557/Drasgow_Hulin_Item_Response_Theory.pdf >.

EMBERTSON, S. E. – REISE, S. P. 2000. *Item Response Theory for Psychologists*. USA: Routledge, 2000. 371 s. ISBN9780805828191

FERRANDO, P.J. – LORENZO-SEVA, U. 2005. IRT-related factor analytics procedures for testing the equivalence of paper-and-pencil and internet-administered questionnaries. *Psychological Methods.* 2005, vol. 10, no. 2, s 193-205.

FUNKHOUSER, G. R. 2001. A Self Anchoring Instrument and Analytical Procedur efor Reducing Cultural Bias in Cross-Cultural Research. *The Journal of Social Psychology*, 2001, vol 133, no. 5, s. 661-673. Dostupný také z WWW: <http://www.wedb.net/download/quanti/mallu_metodos_de_pesquisa/cross%20cultural%20reseatch/artigos%20literatura4/9311297529.pdf>.

GANGLAMAIR-WOOLISCROFT, A. – WOOLISCROFT, B. 2011. A cross-cultural application of the Affective Response to Consumption scale: Investigating US-American and Austrian passengers on long-haul flights. *Journal of Business Research*. 2011. Advance Online Publication. ISSN 0148-2963.

HEND, J. 2009. *Přehled statistických metod : Analýza a metaanalýza dat.* Prah: Portál, 2009. 3. přeprac. vydání. 695 s.

HOOPER, D. - COUGHLAN, J. - MULLEN, M. R. 2008. Structural Equation Modelling : Guidelines for Determining Model Fit." *The Electronic Journal of Business Research Methods*. 2008, vol. 6, no. 1, s. 53 – 60. Dostupný také z WWW: < www.ejbrm.com/issue/download.html?idArticle=183>.

CHYLÍKOVÁ, J. 2011. Úvod do problematiky výzkumu citlivých témat ve výběrových šetřeních. *Sociologický časopis*, 2011, vol. 5, no. 2, s. 185 – 203. Dostupný také z WWW: <http://dav.soc.cas.cz/uploads/a9acd4d331503fbbcb3c6af00345f751b21e1ea7_DaVp185-203%20Chylikova.pdf>.

JANDOUREK, J. 2003. *Úvod do sociologie*. Praha: Portál, 2003. 232 s.

JELÍNEK, M. – KVĚTOŇ, P. – VOBOŘIL, D. 2011. *Testování v psychologii : Teorie odpovědi na položku a počítačové adaptivní testování.* Praha: Grada Publishing, 2011. 158 s. ISBN 9788024735153.

JOHNSON, M. S. 2004. *Item repsonse models and their use in measuring food insecurity and hunger.* Přípěvek na Workshop on the Measurement of Food Insecurity and Hunger, July 15, 2004. Panel to Review USDA's Measurement of Food Insecurity and Hunger [cit. 2012-01-07]. Dostupný z WWW: <http://www7.nationalacademies.org/cnstat/Johnson%20paper.pdf>.

KREIDL, M. 2005. Teorie pro všechny : Metody měření reliability a validity. *Socioweb: Sociologický webzin* [online]. © 2005 [cit. 2012-08-07]. Dostupné z: <http://www.socioweb.cz/index.php?disp=teorie&shw=153&lst=106>.

KREJČÍ, J. 2007. Non-Response in Probability Sample Surveys in the Czech Republic. *Sociologický časopis*, 2007. vol. 43, no. 3, s. 561–587.

KUJAL, P. 2008. *Aplikace teorie odpovědi na položku: Odlišné fungování položek Eysenckova osobnostního dotazníku podle pohlaví*. Diplomová práce. Brno: Masarykova univerzita, 2008. 70 s. Dostupný také z WWW: <http://is.muni.cz/th/75570/ff_m/Diplomova_prace_-_Pavel_Kujal.txt>.

KVĚTON, P - KLIMUSOVÁ, H. 2002. Metodologické aspekty počítačové administrace psychodiagnostických metod. *Československá psychologie.* 2002, vol. 46, no. 3, s. 251-264.

LINN, R. L. 1989. *Has Item Response Theory Increased the Validity of Achievement Test Scores?* Univerity of Colorado: UCLA Center for Research on Evaluation Education, Standards, and Student Testing, 1989. 13 s. Dostupný také z WWW: <.http://www.cse.ucla.edu/products/reports/tr302.pdf> .

LORD, F. N. – NOVICK, M. R. 1968. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley, 1968. 568 s.

MASTERS, G. N. – WRIGHT, B. D. 1997. The Partial Model. In W. J. van der LINDEN – R. K. HAMBLETON (Eds.). *Handbook of Modern Item Response Theory.* New York: Springer, 1997. s 101 – 122.

MASTERS, G. N. – WRIGHT, B. D. 1997. The Partial Model. In W. J. van der LINDEN – R. K. HAMBLETON (Eds.). *Handbook of Modern Item Response Theory.* New York: Springer, 1997. s 101 – 122

MIYAZAKI, K. – HOSHINO, T. 2009. A Bayesian Semiparametric Item Response Model with Dirichlet Process Prior. *Psychometrika*. New York: Springer, 2009. vol. 74, no. 3, s. 375 – 393.

MOLENAAR, I. W. 1997. Nonparametric models for polytomous responses. In W. J. van der LINDEN – R. K. HAMBLETON (Eds.). *Handbook of Modern Item Response Theory.* New York: Springer, 1997. s 369 – 380.

ORLANDO, M. 2004. *Critical Issues to Address when Applying Item Response Theory (IRT) Models.* Bethseda, 2004. 16 s. Příspěvek na konferenci. Dostupný z WWW: <http://edres.org/scripts/cat>.

PADAKI, M. – NATARAJAN, V. 2009. *An Approach to Implementing Adaptive Testing Using Item Response Theory Both Offline and Online*. Minneapolis, 2009. 12 s. Presented at the CAT Research and Applications Around the World Poster Session. Dostupný z WWW: <http://publicdocs.iacat.org/cat2010/cat09natarajan.pdf>.

RASCH, G. B. 1981. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: The University of Chicago Press, 1981. 199 s.

REEVE, B. 2002. *An Introduction to Modern Measurement Theory*. Betsheda: National Cancer Institute, 2002. 67 s. Dostupný také z WWW: <http://appliedresearch.cancer.gov/areas/cognitive/immt.pdf>.

REEVE, B. B. – FAYERS, P. 2005. Applying Item Response Theory Modelling for Evaluating Questionnaire Item and Scale Properties. In P. FAYERS - R. D. HAYS (Eds.). *Assessing Quality of Life in Clinical Trials: Methods of Practice.* 2nd Edition. Oxford University Press, 2005. s. 55-73.

RUDNER, L. M. 1998. *An On-line, Interactive, Computer Adaptive Testing Tutorial* [online]. 1998 [cit. 2012-02-25]. Dostupný z WWW: <http://outcomes.cancer.gov/conference/irt/orlando.pdf>.

SAMEJIMA, F. 1996. Graded Response Model. In W. J. Linden – R. K. Hambleton (Eds). *Handbook of Modern Item Response Theory*. New York: Springer, 1996. s. 85-101.

SCIO. 2008. *Scio* [online]. © 2008 - 2011 [cit. 2012-07-07]. Dostupné z: <http://www.scio.cz/>.

SCIO. 2010. *Personline* [online]. 2010 [cit. 2012-07-07]. Dostupné z: <http://http://www.personline.cz/>.

SHARKNESS, J. – DEANGELO, L. 2011. Measuring Student Involvement: A Comparison of Classical Test Theory in the Construction of Scales from Student Surveys. *Research in Higher Education*, 2011. vol. 52, no. 5, s. 480-507.

STOUT, W. 2005. *DIMTEST (*Version 2.0) [Computer software]. Champaign, IL: The Wiliam Stout Instititute for Measurement.

STUCKY, B. D. 2009. *Item Tesponse Theory for Weighted Summed Scores.* Diuertační práce. The University of North Carolina, 2009. 62 s. Dostupný také z WWW: <http://search.proquest.com/docview/304960217>.

URBÁNEK, T. – ŠIMEČEK, M. 2001. Teorie odpovědi na položku. *Československá psychologie: Časopis pro psychologickou teorii a praxi.* Praha: Academia. 2001, vol. 45, no. 5, s. 428-440. ISSN 0009-062X.

VINOPAL, J. 2008. *Kognitivní přístupy v metodologii výzkumných šetření: metoda okamžité validizace.* Praha: Sociologický ústav AV ČR, v.v.i., 2008. 112 s.

VINOPAL, J. 2011. Indikátor subjektivní kvality pracovního života. *Sociologický časopis*, 2011, vol. 47, no. 5, s. 937 – 965. Dostupný také z WWW: <http://sreview.soc.cas.cz/uploads/c4a04a62ac364e0d1ee63e5de5a3474826cfad83_Vinopal%20soccas2011-5-bezor-3.pdf>.

XU, Y. – IRAN-NEJAD, A. – THOMA, S. J. 2007. Administering Defining Issues Test Online: Do Response Modes Matter? *Journal of Interactive Online Learning.* [online]. Alabama: The University of Alabama. 2007, vol. 6, no. 1., s. 10-27. [cit. 2012-06-29]. ISSN: 1541-4914. Dostupný z WWW <http://www.ncolr.org/jiol/issues/pdf/6.1.2.pdf>.

PŘÍLOHY

Příloha 1

Příloha 1: Baterie 34 měřící subjektivní spokojenost s pracovním životem (Výzkum: Naše společnost 0608, CVVM SOÚ AV ČR, v.v.i.)

POKYN: PŘEDLOŽTE DOTÁZANÉMU KARTU EU.163

EU.163 „Nyní se zaměříme na situaci Vašeho současného zaměstnání a to podle stejných položek, jaké jste hodnotil před chvílí. Tentokrát mi prosím u každé z nich řekněte, jak jste s ní Vy osobně v případě současné hlavní výdělečné činnosti spokojen.

VELMI SPOKOJEN	SPOKOJEN	SPÍŠE SPOKOJEN	SPÍŠE NESPOKOJEN	NESPOKOJEN	VELMI NESPOKOJEN	NETÝKÁ SE	NEVÍ
1	2	3	4	5	6	8	9

a) Jak jste spokojen s výší platu nebo mzdy;	1	2	3	4	5	6	8	9	
b) se spravedlivostí odměňování Vašich pracovních výsledků?	1	2	3	4	5	6	8	9	
c) Jak jste spokojen s nefinančními výhodami plynoucími z Vašeho zaměstnání (např. stravování, delší dovolená, služební auto, telefon apod.)?	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	

d) Jak jste spokojen se vztahy s kolegy;	1	2	3	4	5	6	8	9	
e) s chováním nadřízených;	1	2	3	4	5	6	8	9	
f) s mírou výskytu násilí a šikany na pracovišti?	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	
g) Jak jste spokojen s časovou náročností práce;	1	2	3	4	5	6	8	9	
h) s rozložením pracovní doby?	1	2	3	4	5	6	8	9	
i) Jak jste spokojen s tím, kolik Vám práce umožňuje mít času na Vaši rodinu, zájmy a relaxaci?	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	
j) Jak jste spokojen s tím, jak je Vaše práce zajímavá;	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	
k) s možnostmi dalšího vzdělávání a osobního rozvoje;	1	2	3	4	5	6	8	9	
l) s mírou samostatnosti vykonávané práce?	1	2	3	4	5	6	8	9	
m) Jak jste spokojen s charakterem pracovního poměru, tj. zda jde o poměr hlavní, vedlejší, na dobu neurčitou, na dobu určitou;	1	2	3	4	5	6	8	9	
n) Jak jste spokojen s jistotou pracovního místa?	1	2	3	4	5	6	8	9	
o) Jak jste spokojen s tím, kolik Vám současná práce dává další šance a možnosti uplatnění na trhu práce?	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	
p) Jak jste spokojen s úrovní bezpečnosti práce a ochrany zdraví na pracovišti,	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*9*	
q) s technickým vybavením pracoviště?	1	2	3	4	5	6	8	9	
r) Jak jste spokojen s čistotou, pořádkem a hygienou na pracovišti?“	1	2	3	4	5	6	8	9	

Poznámka: Kurzívou psané otázky označují ty, které je možné vynechat při úspornější verzi modelu – s doménami reprezentovanými pouze dvěma položkami.